CRISPR-Cas: Development and applications for mammalian genome editing

CRISPR-Cas: Development and Applications for Mammalian Genome Editing

Abstract

The ability to introduce targeted modifications into genomes and engineer model organisms holds enormous promise for biomedical and technological applications, and has driven the development of tools such as zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs). To facilitate genome engineering in mammalian cells, we have engineered the CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9 programmable nuclease systems from *Streptococcus pyogenes* SF370 (SpCas9) and *S. thermophilus* LMD-9 (St1Cas9) for mouse and human cell gene editing through heterologous expression of the minimal protein and RNA components. We have demonstrated that Cas9 nucleases can be guided by several short RNAs (sgRNAs) to introduce double stranded breaks (DSB) in the mammalian genome and induce efficient, multiplexed gene modification through non-homologous end-joining-mediated indels or homology-directed repair. Furthermore, we have engineered SpCas9 into a nicking enzyme (SpCas9n) to facilitate recombination while minimizing mutagenic DNA repair processes, and show that SpCas9n can be guided by pairs of appropriately offset sgRNAs to induce DSBs with high efficiency and specificity. In collaboration with Drs. Osamu Nureki and Hiroshi Nishimasu at the University of Tokyo, we further report the crystal

structure of SpCas9 in complex with the sgRNA and target DNA, and elucidate the structure-function relationship of the ribonucleoprotein complex. Finally, through a metagenomic screen of orthologs, we have identified an additional small Cas9 from *Staphylococcus aureus* subsp. *aureus* (SaCas9) that cleaves mammalian endogenous DNA with high efficiency. SaCas9 can be packaged into adeno-associated virus for effective gene modification *in vivo*. Together, these technologies open up exciting possibilities for applications across basic science, biotechnology, and medicine.

Acknowledgements

There are a number of people to whom I'm deeply grateful for guidance, advice, and support.

First and foremost, I'd like to express my gratitude to my advisor, Dr. Feng Zhang, for teaching me how to do science, for leading by example, and for pushing me to always become better. The opportunity to join his lab was a turning point in my training, and the last two years have been some of the most exhilarating times of my life. Feng is an exceptionally creative thinker and equally brilliant teacher, and has continuously supported me in all my endeavors. Under his mentorship, inspired by his infectious enthusiasm, I've learned and grown more than I ever imagined.

I've had the good fortune of working with some of the most passionate and tireless people I've ever known. I'm immensely grateful to all of my co-authors and collaborators, and especially Dr. Le Cong, Patrick Hsu, Dr. Hiroshi Nishimasu, and Winston Yan, who has each taught me much. To all the members of the Zhang lab past and present, and our neighbors in the Xavier and Altshuler labs, thank you for your friendship, help, understanding, and patience. Thank you in particular for the valuable discussion and feedback that has been an integral part of my training. To my former labmates Bernd Zetsche, and Drs. Ronnie Yoo, Juliana Brown, and Guangwen Wang, I'll always treasure the camaraderie and the experiences we've shared.

To Dr. Greg Verdine, I'm deeply indebted for all that you've done for me, for advising me in many things, for your sustained support and faith in me throughout the entirety of my grad school. To Lydia Carmosino, thank you for your kindness and patience, and for your reassurances during many anxious moments. To my dissertation committee, Drs. Andres Leschziner, Fernando Camargo, and Catherine Dulac, thank you for invaluable input, guidance, and assistance.

To my friends in Boston and beyond, I thank you for the joyful times outside of lab, for our travels and adventures, for reminding me to have fun, for keeping in touch even as I disappear into the lab.

Most of all, I'm immensely grateful to my fiancé and long-time friend, Chewie Lin, who has helped grow me so much, personally and professionally, throughout my graduate school years. Thank you for your generosity, your happiness, your critical perspectives, for sharing your passion about science, cooking, and photography, and for your exuberance for life. To my family: you are never too far away from my thoughts, and thank you for all your wisdom, and all the opportunities you've given me.

Table of Contents

List of Abbreviations

Adeno-associated virus	AAV
Cas9 nickase	Cas9n
Clustered regularly interspaced short palindromic repeats	CRISPR
CRISPR-associated gene	Cas
CRISPR-RNA	crRNA
Direct repeat	DR
Double nicking	DN
Double-stranded breaks	DSB
Double-stranded DNA/oligonucleotides	dsDNA/dsODN
Genome-wide off-target	GWOT
Homology-directed repair	HDR
Insertion and deletion	Indel
Non-homologous end-joining	NHEJ
Protospacer-adjacent motif	PAM
Restriction-fragment length polymorphism	RFLP
Single guide RNA	sgRNA
Single-stranded DNA/oligonucleotide	ssDNA/ODN
Small interfering RNAs	siRNA
Staphylococcus aureus Cas9	SaCas9
Streptococcus pyogenes Cas9	SpCas9
Zinc finger (nuclease)	ZF(N)
Transactivating crRNA	tracrRNA
Transcription activator-like effector (nuclease)	TALE(N)
Transcription factor	TF

List of Illustrations

Supplementary Data Figures and Tables

Chapter 1 An introduction to genome engineering

Principles of gene targeting

The understanding that the phenotype of an organism is governed by its genes lies at the foundation of modern biology. Through meticulous observation, selection, and cataloguing, early biologists began to ascribe traits and variations to certain genes in simple model organisms. The need for ways of making desired, precise modifications in the genomes of more complex organisms, then, unsurprisingly, began to drive an ongoing quest for developing ever more versatile and efficient tools for studying both normal biology and disease.

Eukaryotic gene manipulation was first begun over three decades ago in yeast, a system where exogenous gene fragments could be introduced with relative ease; the "gene targeting" technique involved building a vector with a desired modification flanked by arms bearing sequence homology to a given locus, and relying on the native DNA repair pathway of the organism to recombine the modification into the genome(*1-5*). Around the same time, developments in the experimental manipulation of mouse embryos led to the observation that a similar process might occur in mammalian cells(*6, 7*). Over the next several years, a series of landmark papers by Capecchi and

others established the technique of generating transgenic animals by gene targeting in pluripotent mouse embryonic stem (ES) cells via donor template-mediated homologous directed repair (HDR)(*6, 8-11*). Remarkably, leveraging HDR for targeted insertion, deletion, or other modification of genes continues to be one of fundamental tools allowing the generation of countless genetically modified cell lines and organisms.

Powerful as it was, the successful targeting of an endogenous locus is an exceedingly rare event, occurring at the frequency of approximately 1 per 10^5 or more cells(*12, 13*). While in mouse ES cells, a correctly targeted cell line can be selected or screened for, isolated, expanded, and injected into blastocysts to generate chimeric animals, other cell types may not be so amenable to such manipulation. Likewise, any potential of using HDR for *in vivo* gene therapy would be hindered by such frequencies.

What limits the efficiency of HDR? Optimizing the length of homology arms and linearization of the donor plasmid can aid targeting to a modest extent, though surprisingly, increasing the copy number of the donor template seemed to have little effect(*7, 8*). Rather, experiments from flies and yeast yielded the insight that double-stranded breaks (DSB) in the DNA at the target locus can dramatically increase the probability of an HDR event(*14-16*).

Occurring commonly as a result of stalled or collapsed replication forks, DSB in cells are repaired primarily through a couple of pathways (Figure 1-1). The high fidelity pathway typically uses the undamaged sister chromatid as a template, which can be substituted by a donor vector. Alternatively, DSBs can undergo the error-prone non-homologous end joining (NHEJ) process, where broken ends are resected *in situ*. In this way, NHEJ can produce small insertion or deletion (indel) mutations at the site of the break, and less frequently, chromosomal translocations. Indels occurring within the coding region of a gene can in turn result in frameshift mutations that lead to the loss of gene function(*17, 18*).

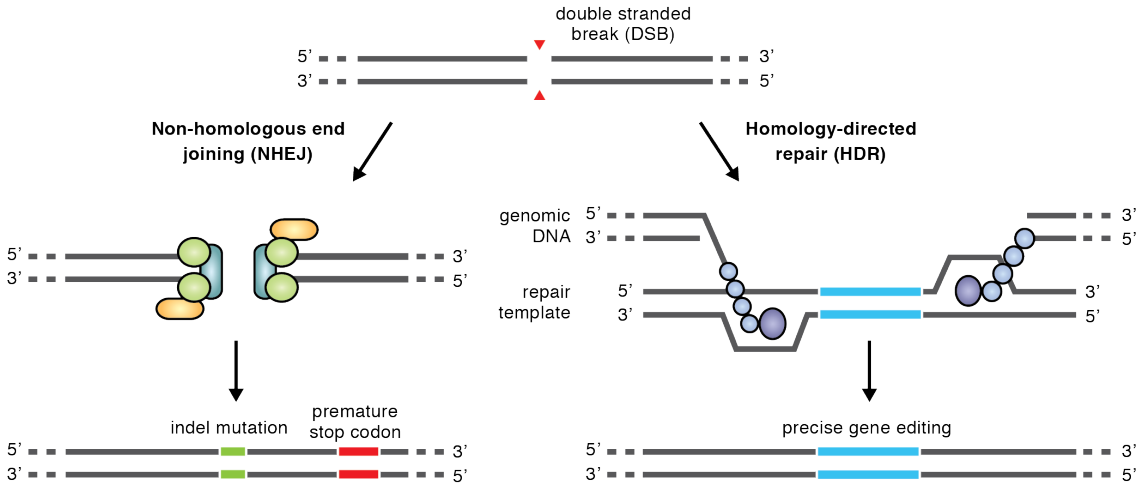


Figure 1-1 DNA double-stranded break repair pathways

The insight that DNA repair mechanisms stimulated by DSBs can be exploited for different types of targeted modification was not lost upon the pioneers of genome engineering. In a simplified conceptual framework, the control of how to effect precision gene editing has become a problem of how to precisely control the timing and location of a DSB. This idea began the development of designed proteins that would allow one to cleave the genome in a controlled manner.

Designer nucleases for genome modification

Taking advantage of the "recombinogenic" nature of DSBs was a family of homing endonucleases already existing naturally in certain unicellular organisms. Such enzymes act similarly as transposons, cleaving DNA in specifically recognized 14 to 40-bp loci and copying (hence homing) its genetic template through damage-induced recombination(*19*). The yeast homing meganuclease I-SceI, so called for its long, rare-cutting 18-bp recognition sequence, became the first such protein to be developed for highly site-specific DNA cleavage for the purpose of gene editing in mammalian cells(*20*).

When Jasin and colleagues transfected I-SceI stably into mouse 3T3 cells, they observed robust cleavage of I-SceI recognition sites; these lesions are then repaired by NHEJ, which could result in deletion of chromosomal

sequences in between two sites. When paired with an HDR donor, I-SceI induced more than 100-fold increase in gene targeting over background(*13*). This result was impressively recapitulated in a transient transfection into mouse ES cells, and subsequently in human cells, fly, zebrafish, and sea anemone among other species, as well as mouse hepatocytes through adenoviral delivery *in vivo*, demonstrating the potential universal applicability of such a system(*12, 20-22*).

Importantly, because their recognition sequences are highly specific, I-SceI appeared to produce little cytotoxicity from off-target cleavage in mammalian cells(*13*). Its exceptional specificity, however, meant on the other hand a handicap for generalized applications, as there are few sites in the mammalian endogenous genome for which the native I-SceI would be useful for targeting. To solve this, several groups sought to reprogram its specificity by re-engineering the enzyme and its homologs (e.g. I-CreI, I-AniI, etc.) using both rational design and combinatorial approaches(*23-26*). Nevertheless, the challenging nature of predictably altering protein recognition spurred the development of alternative technologies.

DNA targeting by ZFNs and TALENs

A key feature, modularity – the ability of each protein subdomain to function independently yet assemble to work together – that would render proteins much more amenable to redesign was missing in homing meganucleases, but not so with zinc fingers (ZF) proteins. Discovered initially in 1985 as a structural motif within a small transcription factor, zinc fingers constitute one of the most abundant DNA recognition modules in metazoans, making up approximately 2% of the human genome(*27-30*). Each finger, approximately 30 amino acids long, specifically contacts 3-bp of DNA in the major groove(*31, 32*); in theory, this means that for a chosen target sequence multiple units of ZFs could be strung together to specifically recognize a longer sequence.

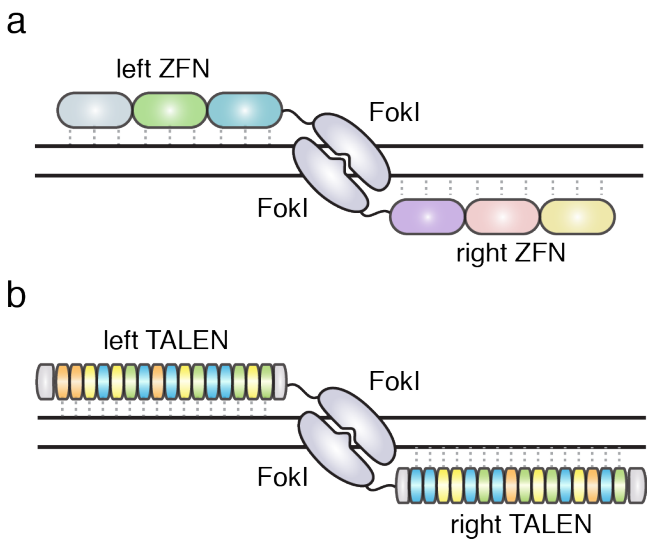


Figure 1-2 Designer nucleases: ZFNs and TALENs

The idea that a fusion protein comprising such a programmable DNA-binding domain and a non-specific nuclease could be created for site-specific cleavage was thus conceived, catalyzing the development of a new generation of synthetic "designer nucleases" (Figure 1-2A). The first demonstration employed a pair of fingers and the catalytic domain of the type II restriction endonuclease *Fok*I, and successfully cleaved predetermined sequences *in vitro* as "chimeric restriction enzymes(*28, 33, 34*)." Others quickly followed by showing the efficacy of zinc finger nucleases (ZFNs) for cleaving endogenous genes and stimulating HDR in *Xenopus* oocytes as well as human embryonic kidney (HEK 293) cells(*35-39*). Importantly, these experiments showed that *Fok*I cleaved as inverted dimers(*35, 37*): that two inversely and appropriately positioned ZFN units directed the cutting of each locus, which could accord it a high degree of specificity since sequence recognition by each half of the ZFN would need to occur together temporally and in defined spatial relationship for cleavage.

Since these groundbreaking demonstrations, a number of developments have enhanced and broadened the use of ZFN for gene targeting. First, finger combinations have been found or designed to recognize many of the 64 triplet nucleotide combinations, though some, such as 5'-GNN-3' triplets, are more reliably recognized(*28*). Related to this, advances in construction of longer ZF arrays allowed recognition of longer stretches of DNA up to

18-bp in length, sufficient for unique targeting of complex genomes(*40, 41*). Secondly, the development of *Fok*I obligate heterodimers(*42, 43*) has ameliorated some of the off-target toxicity arising from unintended homodimerization of the nuclease domains(*44*). With these advances, new methods of efficiently synthesizing ZFNs further paved the way for broader and easier ZFN technology adoption(*45, 46*).

Concomitantly, alternatives to traditional HDR vector designs also expanded the techniques available for precise gene modification. The use of single stranded DNA oligonucleotides (ssDNAs or ssODNs) as donor templates for small changes enabled significant improvements in the efficiency of gene editing and mediating long distance (up to 100-kb) chromosomal microdeletions(*47, 48*). Reminiscent of restriction cloning, donor DNAs with engineered overhangs that match the staggered cleavage patterns generated by ZFNs have been successfully inserted into mammalian chromosomes via NHEJ-based ligation(*49, 50*).

A powerful tool, ZFN has been applied successfully across a multitude of plant and animal species for targeted gene disruption, correction, and addition ranging from single nucleotides to kilobases(*46, 51, 52*). Microinjection of DNA or RNA encoding ZFN targeting endogenous genes into fertilized rat zygotes, for instance, produced animals with 20-100% gene disruptions in a single step; this represents a tremendous advance over traditional ES cell targeting-based transgenic animal generation, not just impressive for the greatly shortened timeframe, but particularly crucial for organisms such as rats where ES cells are not germ line-competent(*53, 54*). ZFNs have thus opened the initial door to genome manipulation of whole organisms, generation of human cell lines for modeling development and disease, as well as a feasible pathway towards gene therapy *in vivo*(*46, 55-57*). Promisingly, one of the early demonstrations of targeted disruption of the *CCR5* gene encoding a co-receptor required for HIV infection is currently in clinical trials, with others sure to follow(*58-61*).

Nonetheless, such a carefully engineered technology is not without its caveats and some significant drawbacks. Early studies of cytotoxicity showed potential lethality in cell lines overexpressing ZFNs(*36, 39*), though improvements in architecture and bioinformatics-based design has ameliorated off-target activity(*42, 43, 46*). In

repairing damaged DNA, HDR and NHEJ pathways can compete and differ in efficiency amongst cell types, thus producing unwanted indels in place of template-based modification. Towards addressing this, ZF-nickases have been engineered by inactivating one of the *Fok*I domains, which reduces mutagenic effects of unwanted DSBs but still allows HDR to occur(*62, 63*). Furthermore, the effect of epigenetics, such as heterochromatin, on ZFN targeting has yet to be fully appreciated.

Perhaps most importantly, ZFs suffer from not being truly independent modular proteins in that the specificities and efficiencies of neighboring modules can influence one another(*64, 65*). Because of this context dependency, the design of ZFNs for new targets isn't straightforward and can often require labor-intensive rounds of screening for optimal ZF domains(*45, 65*). Coupled with difficulty and cost of protein synthesis, the need to pre-validate ZFNs remains a serious barrier for widespread, large-scale adoption. Indeed, much of the ZFN platform used for human endogenous gene editing remains proprietary.

A further step towards both true modularity and universal access in design came with a newer generation of engineered nucleases. Transcription activator-like effectors (TALEs) are proteins secreted by the rice pathogen *Xanthomonas* that are capable of binding promoter sequences and altering the expression of plant genes to aid infection(*66, 67*). The DNA recognition domain of TALEs consists of repetitive monomers that each recognizes a single DNA base pair, which could accord these proteins a much better degree of flexibility in targeting. Within each monomer, typically 33-35 residues in length, DNA recognition is bestowed by two amino acids known as the repeat variable di-residue (RVD), and "cracking the TALE code" became a crucial step in adapting TALEs for genome engineering(*68-70*).

The development of *Fok*I-based TALE-nucleases (TALENs, Figure 1-2B) as a gene editing platform progressed similarly as ZFNs. The modular re-assembly of TALEs and use of TALEN for DNA cleavage was demonstrated first in yeast(*71-73*), followed rapidly in mammalian cells for both NHEJ and HDR applications with efficiencies matching those of ZFNs(*74*). Subsequently, the invariant regions flanking the repeat monomers in the original

TALE scaffold has been further optimized for a simplified architecture(*74, 75*). Like ZFNs, TALENs were swiftly and successfully applied in organisms as diverse as cricket, frog, zebrafish, silkworm, in addition to human somatic and pluripotent stem cells(*76*).

Yet also just like ZFNs or any other technology that can irreversibly modify the genome, questions regarding the specificity of TALENs abounded. In theory, TALENs can achieve a substantial degree of specificity since a minimum of 13-bp of recognition is required for each arm(*77*). Indeed, a few studies examining TALEN targeting of sequences similar to the intended cleavage sites in yeast, human cells, and zebrafish have shown little activity at predicted off-target loci(*73, 75, 78*). Nevertheless, at least two other groups have reported off-target activities(*79, 80*), underlining the need for a genome-wide, unbiased survey of TALEN specificity.

Logistically, the high degree of homology among TALEN monomers presented challenges in protein synthesis and expression. Since PCR-based construction methods are ineffective for such repetitive sequences, the development of alternative cloning strategies became critical(*81-84*). In addition, the possibility of recombination in some situations (*e.g.* lentiviral expression) precludes TALENs from certain applications(*85*). Compounding this, TALENs are much larger than ZFNs per recognized base, and not suitable for delivery via the otherwise promising adeno-associated virus (AAV) vehicle. Finally, though much more modular and predictable, TALEs are not immune from contextual effects(*74*).

More broadly than enabling chimeric nuclease applications, both ZFs and TALEs can provide a generalized platform for customizable DNA binding domains. A number of exciting developments have synthesized this idea with orthogonal functional domains to create sequence-specific ligand-binding proteins, integrases, and histone and DNA modifiers(*77, 86-88*). In this expanded set of tools, some of the most effective and well-characterized technologies are ZF- and TALE-transcriptional modulators that can effect activation or repression of endogenous genes(*41, 74, 82, 89-94*); some of these tools have been additionally engineered to be under chemical or optical control(*88, 95*). Provided the targeting specificity of single ZF or TALE units can be improved to match that of

obligate dimers(*77*), such methods of perturbing endogenous gene expression allows the preservation of natural splicing variants and circumvents the need for delivery of large transgenes, showing promise as therapeutic agents(*86, 96, 97*).

Together, these technologies have already expanded our abilities to alter and modulate the genomes of cell lines and species far beyond those of traditional model systems; they have yielded valuable insights into biological processes, disease modeling, and *in vivo* therapeutics. Where is there room for improvement then? Fundamentally, ZFNs and TALENs share the same principle of relying on decoding and reassembling DNA-binding modules for targeting. Because DNA-protein contacts rely on highly evolved interactions, manipulating their recognition and specificity can be challenging. This is demonstrated by the contextual effects of both ZFs and TALEs, but more generally, the efficacy and specificity of a given ZFN or TALEN pair can be difficult to predict(*45, 65, 81, 83, 98*). Coupled with the need for *de novo* protein synthesis for every construct, the initial validation of targeting can present a hurdle in time and resources and become prohibitive for high-throughput applications. Finally, while both are adept at effecting changes in single genes, the ability to deliver multiple ZFN or TALEN pairs to target several loci (e.g. diseases of polygenic origins or studies on epitasis) presents a significant barrier. All of these reasons lead to one question: are there ways to manipulate genomes using principles wholly different from the existing designer nucleases?

Among the molecular biology techniques used for dissecting functional genetics, one stood out for its potency, versatility, and ease of application. The observation from *C. elegans* that introduction of double-stranded RNAs into cells triggers an antiviral innate immune response to destroy RNA molecules bearing complementary sequences revolutionized our ability to transcriptionally silence genes(*99*). The method is efficient as it's simple: small interfering RNAs (siRNAs) locate their targets by Watson-Crick base-pairing, whose predictability allows the optimization of design parameters for activity and specificity(*100, 101*). Moreover, the RNAi mechanism makes use of conserved endogenous machinery, requiring only the delivery of small RNAs for validation(*102*). This latter point is key to allowing multiplexed and synergistic gene silencing(*103, 104*) as well as creation of

genome-wide siRNA libraries for high throughput screening applications(*105, 106*). For the field of genome engineering, these are enviable qualities, and yet such a system already exists in microbes.

CRISPR-Cas: a programmable RNA-guided endonuclease system

In 1987, a search for the gene encoding a proteolytic enzyme in *E. coli* uncovered a series of regularly spaced repetitive elements containing a "dyad symmetry" or partially palindromic sequences that shared no homology with any known genes(*107*). Even as similar motifs were identified in diverse species of bacteria and archaea, the function of these invariably 21 to 37-bp "direct repeats" and similarly sized non-repetitive "spacers" remained a mystery(*108-111*) and they came to be named clustered regularly interspaced short palindromic repeats (CRISPR) after their most salient structural characteristic(*112*).

The first major breakthrough came about with the increasing availability of bioinformatics data on bacterial and viral genomes. In 2005, three groups independently discovered that bacterial CRISPR spacers shared a high degree of homology with extrachromosomal elements, such as DNA from other species or bacteriophages(*113-115*). Importantly, the spacers seemed to be actively acquired from phages that commonly attacked the bacteria, and those strains that incorporated the spacers into their genomes became resistant to further infection(*113, 115, 116*). It was thus hypothesized that CRISPR served as a kind of "genetic memory of infection," a type of adaptive immune system based on a mechanism then thought to be similar to RNA interference(*117*).

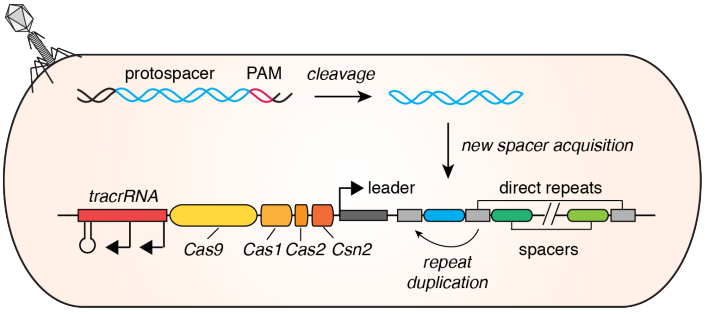


Figure 1-3 Spacer acquisition in the type II CRISPR-Cas adaptive immune system

More thorough analyses revealed that additions of ~30-bp spacers to the CRISPR repeat array always occurred in a polarized manner accompanied by reduplication of repeats(*115, 116, 118*) and additionally required their phage DNA templates or "protospacers" to contain a downstream consensus sequence known as a protospacer adjacent motif (PAM, Figure 1-3)(*119, 120*). For instance, protospacer sequences from phages of the lactobacterium *Streptococcus thermophilus* always immediately preceded a 5'-NNAGAAW-3' PAM sequence, and phages could evolve out of resistance by acquiring mutations in the PAM(*119*). The specific nucleotide composition of the PAM, which is always present in the foreign genetic element but never incorporated into the CRISPR array, vary among the multitude of CRISPR systems(*121*). Though the process of spacer acquisition is not entirely understood, the preclusion of PAMs from the CRISPR array likely plays a role in distinguishing between self vs. non-self(*121-123*).

The mechanism of how target destruction occurred required the identification and understanding of the other components in the CRISPR system. When Schouls and colleagues christened CRISPRs as a new class of repetitive sequences, they also noted that sets of CRISPR-associated (*cas*) protein-coding genes always existed near the repeat arrays, and some of them bore sequence similarity to known helicases and nucleases(*112, 117*). Based on the composition of the *cas* genes, genome architecture of the arrays and their leader sequences, CRISPRs have been subsequently divided into three families that each employ distinct functional mechanisms for spacer incorporation and target cleavage(*124, 125*). Type I systems are defined by a family of proteins that assemble into a multi-subunit CRISPR-associated complex for antiviral defense (Cascade), each of which is essential for a particular function from processing the array transcripts to target degradation. Type III systems are more common in archaea and use additional repeat-associated proteins, with family III-A targeting DNA and III-B single-stranded RNA(*124-127*).

The simplest system, Type II, has thus far been found in bacteria alone and relies solely on its signature protein Cas9 to carry out the adaptation or interference phase(*116, 124, 128*). Here, the CRISPR array is transcribed as a single long mRNA comprising multiple spacers flanked by repeats, which are subsequently processed into discrete

CRISPR RNA (crRNA) units of single spacers followed by direct repeats(*127, 129-131*). A last-to-be-discovered but crucial auxiliary component, transactivating crRNA (tracrRNA), shares partial homology with the crRNA; their duplex assembly, along with Cas9, are required for this co-maturation process(*132*). Mature crRNA:tracrRNA pairings then direct Cas9 proteins to target complementary DNA sequences, with the spacer or guide region of the crRNAs determining specificity by Watson-Crick base pairing (Figure 1-4)(*130*).

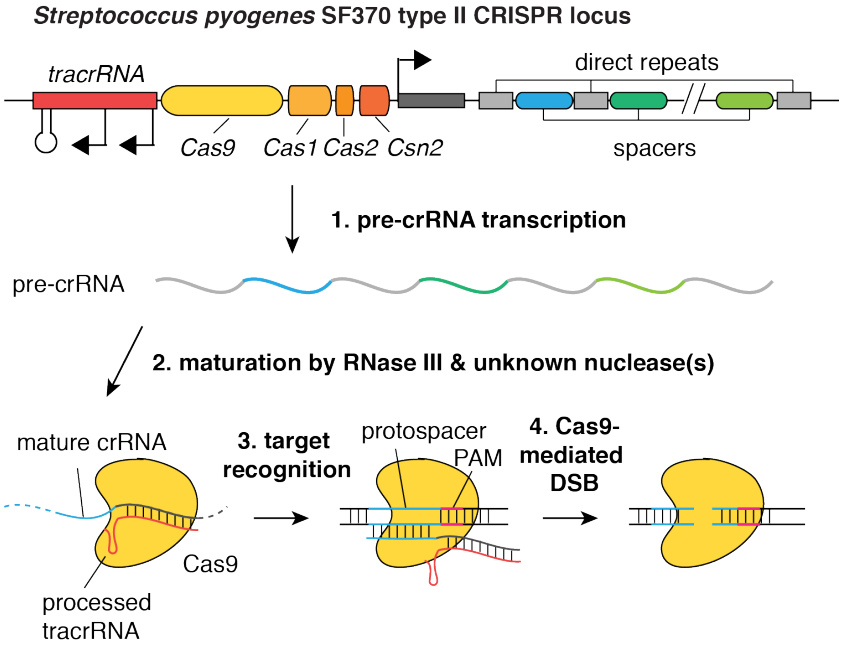


Figure 1-4 Adaptation and target destruction phase in the Type II CRISPR-Cas immune system

As CRISPR systems are identified in more species, sequence analysis reveals extensive horizontal transfer among phylogenetically distant species(*133, 134*). Indeed, the CRISPR locus from *S. thermophilus* could be transplanted into *E. coli* and provide protection against plasmid transformation(*128*). In an *in vitro* cell-free assay, spacer sequences could likewise be altered and allow Cas9 to cleave any sequence, provided the target bears the appropriate PAM. In a further simplification of the native system, crRNA and tracrRNA can be fused via an artificial tetraloop into a chimeric RNA as a single guide RNA (sgRNA) to direct Cas9. Given the sequence homology of Cas9 to other known nucleases of the RuvC and HNH families, catalytic residues on Cas9 have also

been mapped, enabling development of Cas9 nickases as well(*135, 136*). These observations provided the anticipation that CRISPR-Cas could be used for eukaryotic gene editing.

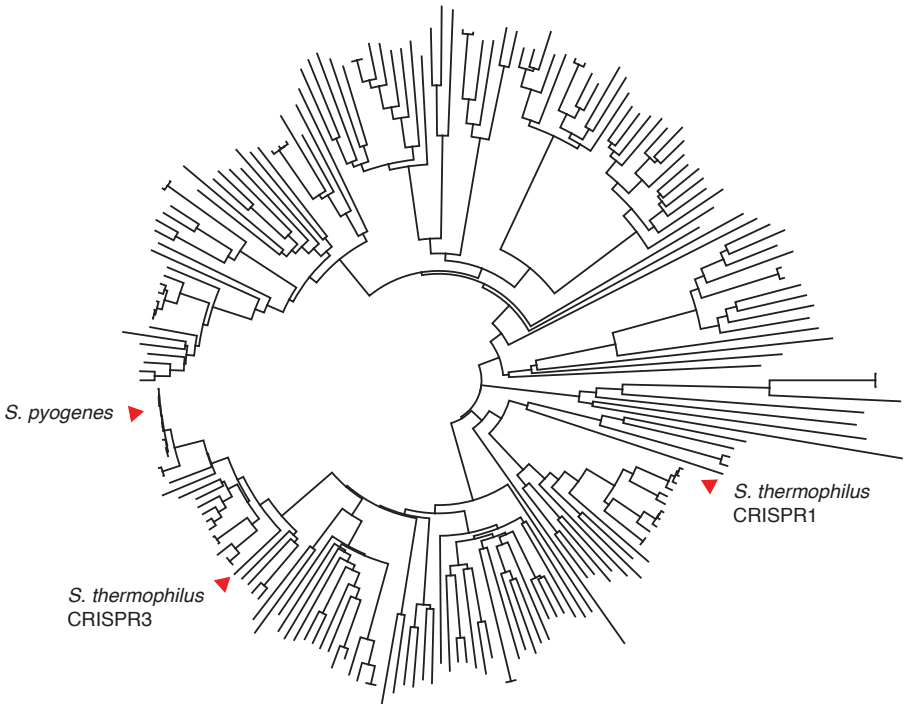


Figure 1-5 Metagenomic diversity of Cas9

In this thesis, I will first describe our efforts at demonstrating the adaptation of Cas9 from *S. pyogenes* and *S. thermophilus* CRISPR1 (SpCas9 and St1Cas9, respectively) for mammalian genome engineering, as we define the minimal set of components needed for a RNA-guided programmable nuclease system(*137*). I will then show the further characterization and optimization of the system for gene disruption and precision editing applications(*138, 139*). As with its predecessors, the specificity of Cas9 raises a significant concern for its suitability for broad and sensitive applications. Many groups, including my colleagues, have attempted to present an accurate and detailed answer to this question(*138, 140-143*). To improve upon SpCas9 specificity, I will describe the development of ZFN- and TALEN-inspired Cas9 double nicking strategy for high fidelity gene editing(*144*). Since solving the crystal structure of Cas9 would prove invaluable in informing further rational mutagenesis and functional alteration of Cas9, I will then present our study in collaboration with Drs. Hiroshi Nishimasu and Osamu Nureki at exploring the structure-function relationship of SpCas9 protein subdomains and their interactions with sgRNA and target DNA(*145*). Finally, I will describe our ongoing efforts at delivering Cas9 for *in vivo* somatic gene editing as we explore the metagenomic diversity of Cas9 (Figure 1-5) for suitable, virally deliverable orthologs.

However, the field of CRISPR-Cas has at the same time been developing with stunning rapidity, and many notable advances have taken place since its first eukaryotic DNA cleavage demonstrations(*137, 146-149*). In technology development, Cas9 has both retraced the path of ZFNs and TALENs as well as inspired novel, inventive applications. These include the development of CRISPR-based transcriptional modulators using catalytically dead Cas9 (dCas9) as well as fusions to natural and synthetic transcription factors while simultaneously taking advantage of the unprecedented multiplex capability of sgRNAs(*88, 89, 142, 150-153*). As a general genome-targeting scaffold, dCas9 has also been adapted for *in vivo* imaging using fluorescence protein fusions(*154, 155*). The ease of oligo synthesis and delivery as well as the modularity of Cas9 targeting has further enabled high-throughput, genome-wide screens to uncover new biology(*156-158*).

In the span of the last year and half, CRISPR has been applied towards *de novo* generation of new transgenic models for a multitude of organisms, including wheat, drosophila, mouse, pig, and marmoset, a truly remarkable

feat(*159-164*). Moreover, demonstrations that Cas9 can be delivered *in vivo* have opened up additional avenues for functional and therapeutic studies(*165, 166*). As we worked to develop double nicking strategy, understand the structure of sgRNA:DNA:SpCas9 complex, and characterize Cas9 orthologs, similar studies have emerged concurrently(*150, 167-170*). The excitement in the field is palpable, surely with more inspiring innovations yet to come for biology, medicine, and technology.

References

1. A. Hinnen, J. Hicks, G. Fink, Transformation of yeast. *Proceedings of the National Academy of Sciences of the United States of America* **75**, 1929 (1978).

2. T. L. Orr-Weaver, J. W. Szostak, R. J. Rothstein, Yeast transformation: a model system for the study of recombination. *Proceedings of the National Academy of Sciences* **78**, (1981).

3. R. J. Rothstein, [12] One-step gene disruption in yeast. *Methods in enzymology* **101**, 202 (1983).

4. N. Rudin, E. Sugarman, J. E. Haber, Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. *Genetics* **122**, 519 (Jul, 1989).

5. A. Plessis, A. Perrin, J. E. Haber, B. Dujon, Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. *Genetics* **130**, 451 (Mar, 1992).

6. K. Folger, K. Thomas, M. R. Capecchi, Analysis of homologous recombination in cultured mammalian cells. *Cold Spring Harbor symposia on quantitative biology* **49**, 123 (1984).

7. M. Capecchi, Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. *Nature reviews. Genetics* **6**, 507 (2005).

8. K. Thomas, K. Folger, M. Capecchi, High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419 (1986).

9. K. Thomas, M. Capecchi, Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell* **51**, 503 (1987).

10. M. Capecchi, Altering the genome by homologous recombination. *Science (New York, N.Y.)* **244**, 1288 (1989).

11. O. Smithies, R. G. Gregg, S. S. Boggs, M. A. Koralewski, R. S. Kucherlapati, Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* **317**, 230 (Sep 19-25, 1985).

12. A. Choulika, A. Perrin, B. Dujon, J. F. Nicolas, Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol Cell Biol* **15**, 1968 (Apr, 1995).

13. P. Rouet, F. Smih, M. Jasin, Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Molecular and cellular biology* **14**, 8096 (1994).

14. F. Paques, J. E. Haber, Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiology and molecular biology reviews : MMBR* **63**, 349 (Jun, 1999).

15. S. C. West *et al.*, Double-strand break repair in human cells. *Cold Spring Harbor symposia on quantitative biology* **65**, 315 (2000).

16. G. B. Gloor, N. A. Nassif, D. M. Johnson-Schlitz, C. R. Preston, W. R. Engels, Targeted gene replacement in Drosophila via P element-induced gap repair. *Science* **253**, 1110 (Sep 6, 1991).

17. D. C. van Gent, J. H. Hoeijmakers, R. Kanaar, Chromosomal stability and the DNA double-stranded break connection. *Nature reviews. Genetics* **2**, 196 (Mar, 2001).

18. S. P. Jackson, J. Bartek, The DNA-damage response in human biology and disease. *Nature* **461**, 1071 (Oct 22, 2009).

19. B. L. Stoddard, Homing endonuclease structure and function. *Quarterly reviews of biophysics* **38**, 49 (Feb, 2005).

20. M. Jasin, Genetic manipulation of genomes with rare-cutting endonucleases. *Trends in genetics : TIG* **12**, 224 (Jun, 1996).

21. F. Smih, P. Rouet, P. Romanienko, M. Jasin, Double-strand breaks at the target locus stimulate gene targeting in embryonic stem cells. *Nucleic acids research* **23**, 5012 (1995).

22. G. Silva *et al.*, Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Current gene therapy* **11**, 11 (2011).

23. J. Smith *et al.*, A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic acids research* **34**, e149 (2006).

24. B. S. Chevalier *et al.*, Design, activity, and structure of a highly specific artificial endonuclease. *Mol Cell* **10**, 895 (Oct, 2002).

25. J. C. Epinat *et al.*, A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic acids research* **31**, 2952 (Jun 1, 2003).

26. D. Sussman *et al.*, Isolation and characterization of new homing endonuclease specificities at individual target site positions. *Journal of molecular biology* **342**, 31 (Sep 3, 2004).

27. D. Jantz, B. Amann, G. Gatto, J. Berg, The design of functional DNA-binding proteins based on zinc finger domains. *Chemical reviews* **104**, 789 (2004).

28. M. Porteus, D. Carroll, Gene targeting using zinc finger nucleases. *Nature biotechnology* **23**, 967 (2005).

29. J. Miller, A. McLachlan, A. Klug, Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal* **4**, 1609 (1985).

30. G. Diakun, L. Fairall, A. Klug, EXAFS study of the zinc-binding sites in the protein transcription factor IIIA. *Nature* **324**, 698 (1986).

31. N. P. Pavletich, C. O. Pabo, Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252**, 809 (May 10, 1991).

32. S. A. Wolfe, L. Nekludova, C. O. Pabo, DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure* **29**, 183 (2000).

33. S. Chandrasegaran, J. Smith, Chimeric restriction enzymes: what is next? *Biological chemistry* **380**, 841 (Jul-Aug, 1999).

34. Y. Kim, J. Cha, S. Chandrasegaran, Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 1156 (1996).

35. M. Bibikova *et al.*, Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol Cell Biol* **21**, 289 (Jan, 2001).

36. M. Bibikova, M. Golic, K. Golic, D. Carroll, Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* **161**, 1169 (2002).

37. J. Smith *et al.*, Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. *Nucleic acids research* **28**, 3361 (2000).

38. M. Bibikova, K. Beumer, J. Trautman, D. Carroll, Enhancing gene targeting with designed zinc finger nucleases. *Science (New York, N.Y.)* **300**, 764 (2003).

39. M. H. Porteus, D. Baltimore, Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (May 2, 2003).

40. Q. Liu, D. Segal, J. Ghiara, C. Barbas, Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 5525 (1997).

41. R. Beerli, C. Barbas, Engineering polydactyl zinc-finger transcription factors. *Nature biotechnology* **20**, 135 (2002).

42. J. Miller *et al.*, An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**, 778 (2007).

43. M. Szczepek *et al.*, Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nature biotechnology* **25**, 786 (2007).

44. K. Beumer, G. Bhattacharyya, M. Bibikova, J. Trautman, D. Carroll, Efficient gene targeting in Drosophila with zinc-finger nucleases. *Genetics* **172**, 2391 (2006).

45. T. Cathomen, J. Joung, Zinc-finger nucleases: the next generation emerges. *Molecular therapy : the journal of the American Society of Gene Therapy* **16**, 1200 (2008).

46. F. Urnov, E. Rebar, M. Holmes, H. Zhang, P. Gregory, Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics* **11**, 636 (2010).

47. F. Chen *et al.*, High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nature methods* **8**, 753 (2011).

48. H. Lee, E. Kim, J.-S. Kim, Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome research* **20**, 81 (2010).

49. S. Orlando *et al.*, Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic acids research* **38**, (2010).

50. M. Maresca, V. Lin, N. Guo, Y. Yang, Obligate ligation-gated recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome research* **23**, 539 (2013).

51. T. Gaj, C. Gersbach, C. Barbas, ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology*, (2013).

52. E. Moehle *et al.*, Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 3055 (2007).

53. T. Mashimo *et al.*, Generation of knockout rats with X-linked severe combined immunodeficiency (X-SCID) using zinc-finger nucleases. *PloS one* **5**, (2010).

54. A. Geurts *et al.*, Knockout rats via embryo microinjection of zinc-finger nucleases. *Science (New York, N.Y.)* **325**, 433 (2009).

55. A. Lombardo *et al.*, Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nature biotechnology* **25**, 1298 (2007).

56. F. Urnov *et al.*, Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646 (2005).

57. F. Soldner *et al.*, Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* **146**, 318 (2011).

58. E. Perez *et al.*, Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**, 808 (2008).

59. T. Pablo *et al.*, Gene Editing of CCR5 in Autologous CD4 T Cells of Persons Infected with HIV. *New England Journal of Medicine* **370**, (2014).

60. V. Sebastiano *et al.*, In situ genetic correction of the sickle cell anemia mutation in human induced pluripotent stem cells using engineered zinc finger nucleases. *Stem cells (Dayton, Ohio)* **29**, 1717 (2011).

61. K. Yusa *et al.*, Targeted gene correction of α1-antitrypsin deficiency in induced pluripotent stem cells. *Nature* **478**, 391 (2011).

62. C. Ramirez *et al.*, Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. *Nucleic acids research* **40**, 5560 (2012).

63. E. Kim *et al.*, Precision genome engineering with programmable DNA-nicking enzymes. *Genome research* **22**, 1327 (2012).

64. J. Miller, C. Pabo, Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *Journal of molecular biology* **313**, 309 (2001).

65. J. Sander *et al.*, Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nature methods* **8**, 67 (2011).

66. K. Gu *et al.*, R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature* **435**, 1122 (2005).

67. S. Kay, S. Hahn, E. Marois, G. Hause, U. Bonas, A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science (New York, N.Y.)* **318**, 648 (2007).

68. D. Deng *et al.*, Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science (New York, N.Y.)* **335**, 720 (2012).

69. M. Moscou, A. Bogdanove, A simple cipher governs DNA recognition by TAL effectors. *Science (New York, N.Y.)* **326**, 1501 (2009).

70. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)* **326**, 1509 (2009).

71. M. Christian, T. Cermak, E. Doyle, C. Schmidt …, Targeting DNA double-strand breaks with TAL effector nucleases. …, (2010).

72. T. Li *et al.*, TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic acids research* **39**, 359 (2011).

73. T. Li, S. Huang, X. Zhao, D. Wright…, Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic acids …*, (2011).

74. J. Miller *et al.*, A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**, 143 (2011).

75. C. Mussolino, R. Morbitzer, F. Lütge ... , A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids ...* , (2011).

76. J. Joung, J. Sander, TALENs: a widely applicable technology for targeted genome editing. *Nature reviews. Molecular cell biology* **14**, 49 (2013).

77. A. Bogdanove, D. Voytas, TAL effectors: customizable proteins for DNA targeting. *Science (New York, N.Y.)* **333**, 1843 (2011).

78. P. Huang *et al.*, Heritable gene targeting in zebrafish using customized TALENs. *Nature biotechnology* **29**, 699 (Aug, 2011).

79. L. Tesson *et al.*, Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* **29**, 695 (2011).

80. D. Hockemeyer *et al.*, Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**, 731 (2011).

81. T. Cermak *et al.*, Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **39**, (2011).

82. F. Zhang *et al.*, Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**, 149 (2011).

83. D. Reyon *et al.*, FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology* **30**, 460 (2012).

84. N. E. Sanjana *et al.*, A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* **7**, 171 (Jan, 2012).

85. M. Holkers *et al.*, Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic acids research* **41**, e63 (Mar 1, 2013).

86. A. Jamieson, J. Miller, C. Pabo, Drug discovery with engineered zinc-finger proteins. *Nature reviews. Drug discovery* **2**, 361 (2003).

87. M. Maeder *et al.*, Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nature biotechnology* **31**, 1137 (2013).

88. S. Konermann *et al.*, Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**, 472 (Aug 22, 2013).

89. M. L. Maeder *et al.*, CRISPR RNA-guided activation of endogenous human genes. *Nature methods*, (Jul 25, 2013).

90. R. Geissler *et al.*, Transcriptional activators of human genes with programmable DNA-specificity. *PloS one* **6**, (2011).

91. L. Cong, R. Zhou, Y.-C. Kuo, M. Cunniff, F. Zhang, Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications* **3**, 968 (2012).

92. R. Beerli, B. Dreier, C. Barbas, Positive and negative regulation of endogenous genes by designed transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 1495 (2000).

93. G. Cuthbert *et al.*, Histone deimination antagonizes arginine methylation. *Cell* **118**, 545 (2004).

94. P. Perez-Pinera *et al.*, RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature methods*, (Jul 25, 2013).

95. L. Polstein, C. Gersbach, Light-inducible spatiotemporal control of gene activation by customizable zinc finger transcription factors. *Journal of the American Chemical Society* **134**, 16480 (2012).

96. E. Rebar *et al.*, Induction of angiogenesis in a mouse model using engineered transcription factors. *Nature medicine* **8**, 1427 (2002).

97. A. Reik *et al.*, Enhanced protein production by engineered zinc finger proteins. *Biotechnology and bioengineering* **97**, 1180 (2007).

98. C. Ramirez *et al.*, Unexpected failure rates for modular assembly of engineered zinc fingers. *Nature methods* **5**, 374 (2008).

99. A. Fire *et al.*, Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806 (1998).

100. J.-P. Vert, N. Foveau, C. Lajaunie, Y. Vandenbrouck, An accurate and interpretable model for siRNA efficacy prediction. *BMC bioinformatics* **7**, 520 (2006).

101. M. Amarzguioui, H. Prydz, An algorithm for selection of functional siRNA sequences. *Biochemical and biophysical research communications* **316**, 1050 (2004).

102. E. Bernstein, A. Caudy, S. Hammond, G. Hannon, Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363 (2001).

103. D. Gou *et al.*, A novel approach for the construction of multiple shRNA expression vectors. *The journal of gene medicine* **9**, 751 (2007).

104. L. Lambeth, N. Van Hateren, S. Wilson, V. Nair, A direct comparison of strategies for combinatorial RNA interference. *BMC molecular biology* **11**, 77 (2010).

105. J. Silva *et al.*, Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science (New York, N.Y.)* **319**, 617 (2008).

106. J. Moffat *et al.*, A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283 (2006).

107. Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, A. Nakata, Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of bacteriology* **169**, 5429 (1987).

108. F. Mojica, C. Díez-Villaseñor, E. Soria, G. Juez, Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular microbiology* **36**, 244 (2000).

109. A. van Belkum, S. Scherer, L. van Alphen, H. Verbrugh, Short-sequence DNA repeats in prokaryotic genomes. *Microbiology and molecular biology reviews : MMBR* **62**, 275 (1998).

110. F. Mojica, C. Ferrer, G. Juez, F. Rodríguez-Valera, Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Molecular microbiology* **17**, 85 (1995).

111. N. Hoe *et al.*, Rapid molecular genetic subtyping of serotype M1 group A Streptococcus strains. *Emerging infectious diseases* **5**, 254 (1999).

112. R. Jansen, J. Embden, W. Gaastra, L. Schouls, Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology* **43**, 1565 (2002).

113. A. Bolotin, B. Quinquis, A. Sorokin, S. Ehrlich, Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology (Reading, England)* **151**, 2551 (2005).

114. F. Mojica, C. Díez-Villaseñor, J. García-Martínez, E. Soria, Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution* **60**, 174 (2005).

115. C. Pourcel, G. Salvignol, G. Vergnaud, CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology (Reading, England)* **151**, 653 (2005).

116. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)* **315**, 1709 (2007).

117. K. Makarova, N. Grishin, S. Shabalina, Y. Wolf, E. Koonin, A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology direct* **1**, 7 (2006).

118. J. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67 (2010).

119. H. Deveau *et al.*, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of bacteriology* **190**, 1390 (2008).

120. P. Horvath *et al.*, Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol* **190**, 1401 (Feb, 2008).

121. F. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, England)* **155**, 733 (2009).

122. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea. *Science (New York, N.Y.)* **327**, 167 (2010).

123. L. Marraffini, E. Sontheimer, Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**, 568 (2010).

124. K. Makarova *et al.*, Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology* **9**, 467 (2011).

125. B. Wiedenheft, S. Sternberg, J. Doudna, RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331 (2012).

126. T. Sinkunas *et al.*, Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* **30**, 1335 (Apr 6, 2011).

127. C. Hale *et al.*, RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945 (2009).

128. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39**, 9275 (2011).

129. T.-H. Tang *et al.*, Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Molecular microbiology* **55**, 469 (2005).

130. S. J. Brouns *et al.*, Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science (New York, N.Y.)* **321**, 960 (2008).

131. R. Lillestøl *et al.*, CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Molecular microbiology* **72**, 259 (2009).

132. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602 (2011).

133. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167 (Jan 8, 2010).

134. J. Godde, A. Bickerton, The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *Journal of molecular evolution* **62**, 718 (2006).

135. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816 (2012).

136. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 86 (2012).

137. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

138. P. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827 (2013).

139. F. Ran *et al.*, Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281 (2013).

140. Y. Fu *et al.*, High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature biotechnology* **31**, 822 (Sep, 2013).

141. V. Pattanayak *et al.*, High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature biotechnology* **31**, 839 (Sep, 2013).

142. M. H. Larson *et al.*, CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc* **8**, 2180 (Nov, 2013).

143. D. Bikard *et al.*, Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic acids research* **41**, 7429 (2013).

144. F. Ran *et al.*, Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380 (2013).

145. H. Nishimasu *et al.*, Crystal structure of cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935 (2014).

146. P. Mali *et al.*, RNA-guided human genome engineering via Cas9. *Science* **339**, 823 (Feb 15, 2013).

147. S. W. Cho, S. Kim, J. M. Kim, J. S. Kim, Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology* **31**, 230 (Mar, 2013).

148. W. Hwang *et al.*, Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology* **31**, 227 (2013).

149. M. Jinek *et al.*, RNA-programmed genome editing in human cells. *eLife* **2**, (2013).

150. P. Mali *et al.*, CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**, 833 (2013).

151. F. Farzadfard, S. D. Perli, T. K. Lu, Tunable and Multifunctional Eukaryotic Transcription Factors Based on CRISPR/Cas. *ACS synthetic biology*, (Sep 11, 2013).

152. A. W. Cheng *et al.*, Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell research* **23**, 1163 (Oct, 2013).

153. L. S. Qi *et al.*, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173 (Feb 28, 2013).

154. B. Chen *et al.*, Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479 (Dec 19, 2013).

155. T. Anton, S. Bultmann, H. Leonhardt, Y. Markaki, Visualization of specific DNA sequences in living mouse embryonic stem cells with a programmable fluorescent CRISPR/Cas system. *Nucleus* **5**, (Mar 12, 2014).

156. O. Shalem *et al.*, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84 (Jan 3, 2014).

157. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80 (Jan 3, 2014).

158. H. Koike-Yusa, Y. Li, E. P. Tan, M. D. Velasco-Herrera, K. Yusa, Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature biotechnology*, (Dec 23, 2013).

159. Y. Niu *et al.*, Generation of Gene-Modified Cynomolgus Monkey via Cas9/RNA-Mediated Gene Targeting in One-Cell Embryos. *Cell* **156**, 836 (Feb 13, 2014).

160. H. Wang *et al.*, One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910 (2013).

161. Z. Yu *et al.*, Highly Efficient Genome Modifications Mediated by CRISPR/Cas9 in Drosophila. *Genetics*, (Jul 5, 2013).

162. T. Hai, F. Teng, R. Guo, W. Li, Q. Zhou, One-step generation of knockout pigs by zygote injection of CRISPR/Cas system. *Cell research* **24**, 372 (Mar, 2014).

163. Q. Shan *et al.*, Targeted genome modification of crop plants using a CRISPR-Cas system. *Nature biotechnology* **31**, 686 (Aug, 2013).

164. D. Mashiko *et al.*, Feasibility for a large scale mouse mutagenesis by injecting CRISPR/Cas plasmid into zygotes. *Development, growth & differentiation* **56**, 122 (Jan, 2014).

165. S. Ramakrishna *et al.*, Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res*, (Apr 2, 2014).

166. H. Yin *et al.*, Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. *Nature biotechnology*, (Mar 30, 2014).

167. I. Fonfara *et al.*, Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic acids research* **42**, 2577 (Feb 1, 2014).

168. Z. Hou *et al.*, Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15644 (Sep 24, 2013).

169. K. M. Esvelt *et al.*, Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature methods* **10**, 1116 (Nov, 2013).

170. M. Jinek *et al.*, Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (Mar 14, 2014).

Chapter 2 Development of CRISPR-Cas for mammalian genome editing

Le Cong[1,4,*], F. Ann Ran[1,6,*], Patrick D. Hsu[1,6], Xuebing Wu[3], Shuailiang Lin[1,7], Jason Wright[1], David Cox[1,5], Robert Barretto[8], and Feng Zhang[1,2]

[1]Broad Institute and [2]McGovern Institute
[3]Computational and Systems Biology Graduate Program
and Koch Institute for Integrative Cancer Research
Massachusetts Institute of Technology, Cambridge, MA 02139 USA

[4]Program in Biological and Biomedical Sciences
[5]Harvard-MIT Health Sciences and Technology
Harvard Medical School, Boston, MA 02115 USA

[6]Department of Molecular and Cellular Biology
Harvard University, Cambridge, MA 02138 USA

[7]School of Life Sciences,
Tsinghua University, Beijing, 100084, China

[8]Department of Biochemistry and Molecular Biophysics
College of Physicians and Surgeons,
Columbia University, New York, NY 10065, USA
*These authors contributed equally to this work.

Acknowledgements

We thank the entire Zhang lab for their support and advice; P. A. Sharp for help with Northern blot analysis; and X. Ye for help with confocal imaging. This work was supported by a NIH Director's Pioneer Award (DP1-MH100706) to F.Z., the Keck, McKnight, Gates, Damon Runyon, Searle Scholars, Klingenstein, and Simons Foundations, Bob Metcalfe, Mike Boylan, and Jane Pauley.

Author contribution

L.C. and F.Z. conceived the study, L.C., F.A.R., P.D.H., and F.Z. designed and performed the experiments with assistance from all authors. F.Z. supervised all aspects of the study. F.A.R. and F.Z. wrote the manuscripts with help from all authors. In particular, I carried out the immunofluorescence experiments, TALEN and Cas9 comparison, and devised and conducted the multiplex targeting, ssODN-mediated homologous recombination, and *in vitro* methylation experiments, and designed and cloned the sgRNA optimization vectors.

This chapter contains work from the manuscripts "Multiplex Genome Engineering using CRISPR/Cas Systems," published in *Science*, February 15, 2013, Vol. 339 no. 6121 pp. 819-821; and "Genome engineering using the CRISPR-Cas9 system," published in *Nature Protocols*, October 13, 2013, Vol. 8 no. 11 pp. 2281-2308, and "DNA targeting specificity of RNA-guided Cas9 nucleases," published in *Nature Biotechnology*, July 21 2013, Vol. 31 no. 9 pp. 827-832. The text and figures were modified to fit the format of this dissertation.

Introduction

Precise and efficient genome targeting technologies are needed to enable systematic reverse engineering of causal genetic variations by allowing selective perturbation of individual genetic elements. Although genome-editing technologies such as designer zinc fingers (ZFs) (*1-4*), transcription activator-like effectors (TALEs) (*4-10*), and homing meganucleases (*12*) have begun to enable targeted genome modifications, there remains a need for new technologies that are scalable, affordable, and easy to engineer. Here, we report the development of a new class of precision genome engineering tools based on the RNA-guided Cas9 nuclease (*13-15*) from the type II prokaryotic CRISPR adaptive immune system(*16-19*).

The *Streptococcus pyogenes* SF370 type II CRISPR locus consists of four genes, including the Cas9 nuclease, as well as two non-coding RNAs: tracrRNA and a pre-crRNA array containing nuclease guide sequences (spacers) interspaced by identical direct repeats (DRs) (*20*). We sought to harness this prokaryotic RNA-programmable nuclease system to introduce targeted double stranded breaks (DSBs) in mammalian chromosomes through heterologous expression of the key components. It has been previously shown that expression of tracrRNA, pre-crRNA, host factor RNase III, and Cas9 nuclease are necessary and sufficient for cleavage of DNA *in vitro* (*13, 14*) and in prokaryotic cells (*21, 22*).

Adaptation of CRISPR-Cas for multiplexed mammalian gene editing

We codon optimized the *S. pyogenes Cas9 (SpCas9)* and *RNase III (SpRNase III)* and attached nuclear localization signals (NLS) to ensure nuclear compartmentalization in mammalian cells. Expression of these constructs in human 293FT cells revealed that two NLSs are most efficient at targeting SpCas9 to the nucleus (Figure 2-1A). To reconstitute the non-coding RNA components of CRISPR, we expressed an 89-nucleotide (nt) tracrRNA

(Supplementary Figure 1A) under the RNA polymerase III U6 promoter (Figure 2-1B). Similarly, we used the U6 promoter to drive the expression of a pre-crRNA array comprising a single guide spacer flanked by DRs (Figure 2-1B). We designed our initial spacer to target a 30-basepair (bp) site (protospacer) in the human *EMX1* locus that precedes an NGG, the requisite protospacer adjacent motif (PAM) (Figure 2-1C) (*23, 24*).

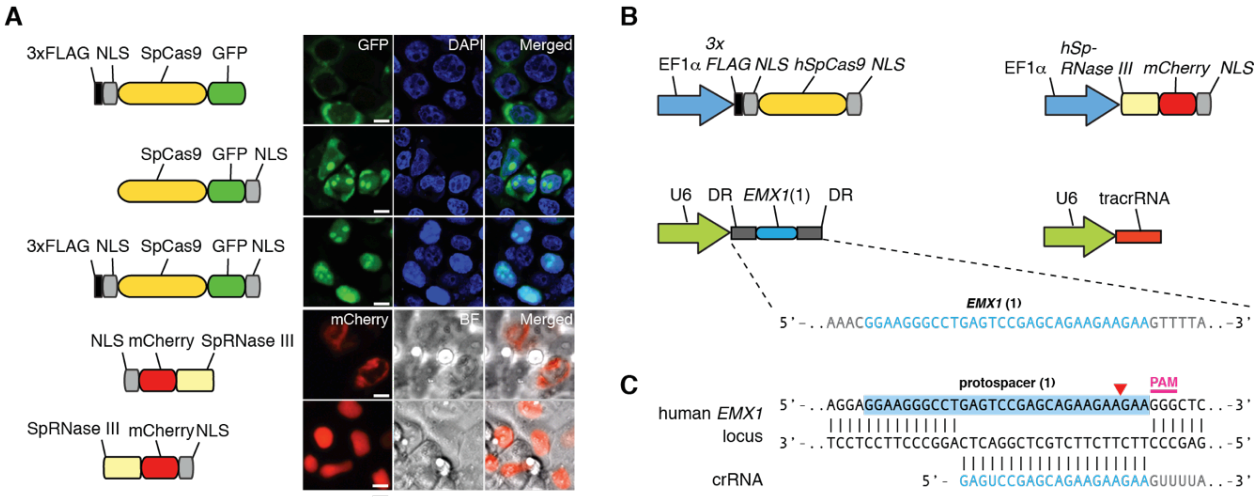


Figure 2-1 Reconstitution of the Type II CRISPR locus from Streptococcus pyogenes SF370 for mammalian gene targeting.

(A) Engineering of SpCas9 and SpRNase III with NLSs enables import into the mammalian nucleus. (B) Mammalian expression of SpCas9 and SpRNase III are driven by the EF1a promoter, whereas tracrRNA and pre-crRNA array (DR-Spacer-DR) are driven by the U6 promoter. A protospacer (blue highlight) from the human *EMX1* locus with PAM is used as template for the spacer in the pre-crRNA array. (C) Schematic representation of base pairing between target locus and *EMX1*-targeting crRNA. Red arrow indicates putative cleavage site.

To test whether heterologous expression of the CRISPR system (SpCas9, SpRNase III, tracrRNA, and pre-crRNA) can achieve targeted cleavage of mammalian chromosomes, we transfected 293FT cells with different combinations of CRISPR components. Since DSBs in mammalian DNA are partially repaired by the indel-forming non-homologous end joining (NHEJ) pathway, we used the SURVEYOR assay (Supplementary Figure 2) to detect endogenous target cleavage (Figure 2-2A and Supplementary Figure 1B). Co-transfection of all four required CRISPR components resulted in efficient cleavage of the protospacer (Figure 2-2A and Supplementary Figure 1B), which is subsequently verified by Sanger sequencing (Figure 2-2B). Interestingly, SpRNase III was

not necessary for cleavage of the protospacer (Figure 2-2A), and the 89-nt tracrRNA is processed in its absence (Supplementary Figure 1C). Similarly, maturation of pre-crRNA does not require RNase III (Figure 2-2A and Supplementary Figure 3), suggesting that there may be endogenous mammalian RNases that assist in pre-crRNA maturation (*25-27*). Removing any of the remaining RNA or Cas9 components abolished the genome cleavage activity of the CRISPR system (Figure 2-2A). These results define a minimal three-component system for efficient CRISPR-mediated genome modification in mammalian cells.

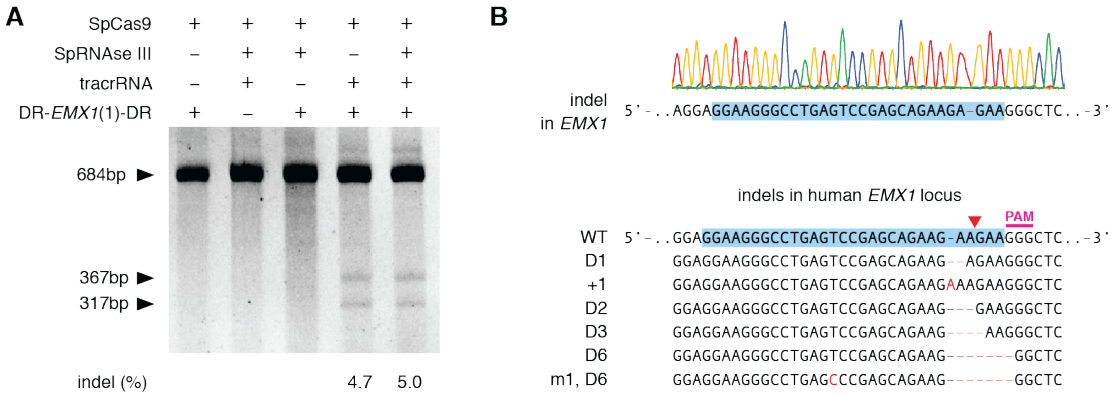


Figure 2-2 SpCas9-mediated indels

(A) SURVEYOR assay for SpCas9-mediated indels. (B) An example chromatogram showing a micro-deletion, as well as representative sequences of mutated alleles identified from 187 clonal amplicons. Red dashes, deleted bases; red bases, insertions or mutations. Scale bar = 10μm.

Next, we explored the generalizability of CRISPR-mediated cleavage in eukaryotic cells by targeting additional protospacers within the *EMX1* locus (Figure 2-3A). To improve co-delivery, we designed an expression vector to drive both pre-crRNA and SpCas9 (Supplementary Figure 4). In parallel, we adapted a chimeric crRNA-tracrRNA hybrid (Figure 2-3B, top) design recently validated *in vitro* (*13*), where a mature crRNA is fused to a partial tracrRNA via a synthetic stem-loop to mimic the natural crRNA:tracrRNA duplex (Figure 2-3B, bottom). We observed cleavage of all protospacer targets when SpCas9 is co-expressed with pre-crRNA (DR-spacer-DR)

and tracrRNA. However, not all chimeric RNA designs could facilitate cleavage of their genomic targets (Figure 2-3C, Supplementary Table 1). We then tested targeting of additional genomic loci in both human and mouse cells by designing pre-crRNAs and chimeric RNAs targeting the human *PVALB* and the mouse *Th* loci (Supplementary Figure 5). We achieved efficient modification at all three mouse *Th* and one *PVALB* targets using the crRNA:tracrRNA design, thus demonstrating the broad applicability of the CRISPR system in modifying different loci across multiple organisms (Supplementary Table 1). For the same protospacer targets, cleavage efficiencies of chimeric RNAs were either lower than those of crRNA:tracrRNA duplexes or undetectable. This may be due to differences in the expression and stability of RNAs, degradation by endogenous RNAi machinery, or secondary structures leading to inefficient Cas9 loading or target recognition.

Effective genome editing requires that nucleases target specific genomic loci with both high precision and efficiency. To investigate the specificity of CRISPR-mediated cleavage, we analyzed single-nucleotide mismatches between the spacer and its mammalian protospacer target (Figure 2-4A). We observed that single-base mismatch up to 12-bp 5' of the PAM completely abolished genomic cleavage by SpCas9, whereas spacers with mutations farther upstream retained activity against the protospacer target (Figure 2-4B). This is consistent with previous bacterial and *in vitro* studies of Cas9 specificity (*13, 21*). Furthermore, CRISPR is able to mediate genomic cleavage as efficiently as a pair of TALE nucleases (TALEN) targeting the same *EMX1* protospacer (Figure 2-5).

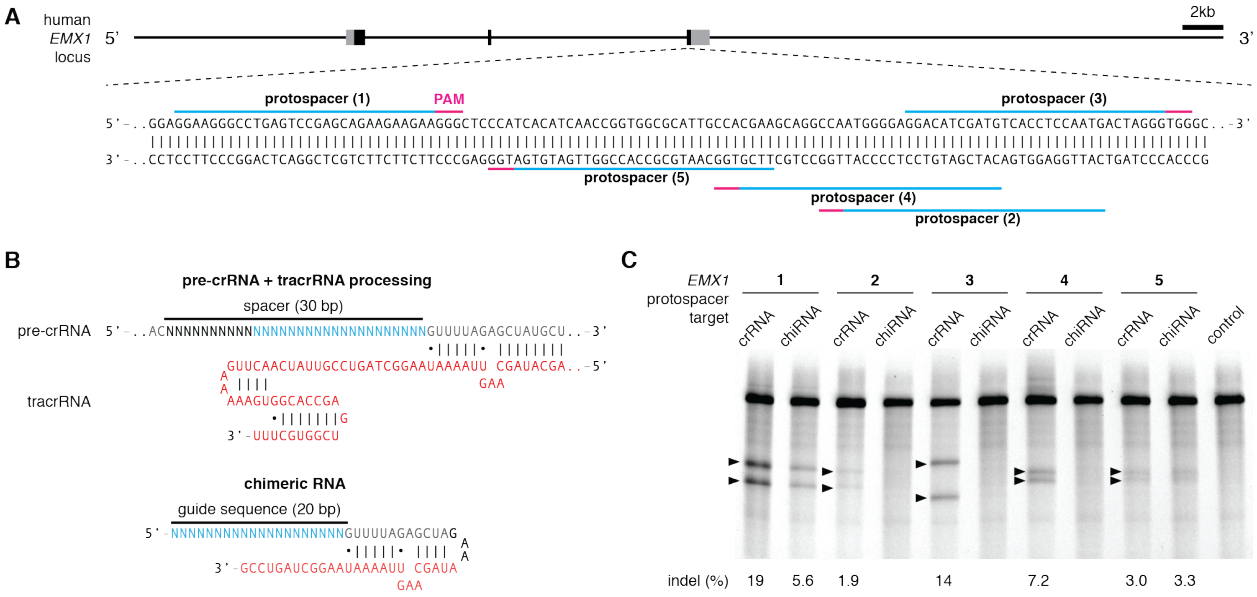


Figure 2-3 SpCas9 can be reprogrammed to target multiple genomic loci in mammalian cells.

(A) Schematic of the human *EMX1* locus showing the location of five protospacers, indicated by blue lines with corresponding PAM in magenta. (B) Schematic of the pre-crRNA:tracrRNA complex (top) showing hybridization between the direct repeat (gray) region of the pre-crRNA and tracrRNA. Schematic of a chimeric RNA design (*12*) (bottom). tracrRNA sequence is shown in red and the 20bp spacer sequence in blue. (C) SURVEYOR assay comparing the efficacy of Cas9-mediated cleavage at five protospacers in the human *EMX1* locus. Each protospacer is targeted using either processed pre-crRNA:tracrRNA complex (crRNA) or chimeric RNA (chiRNA).

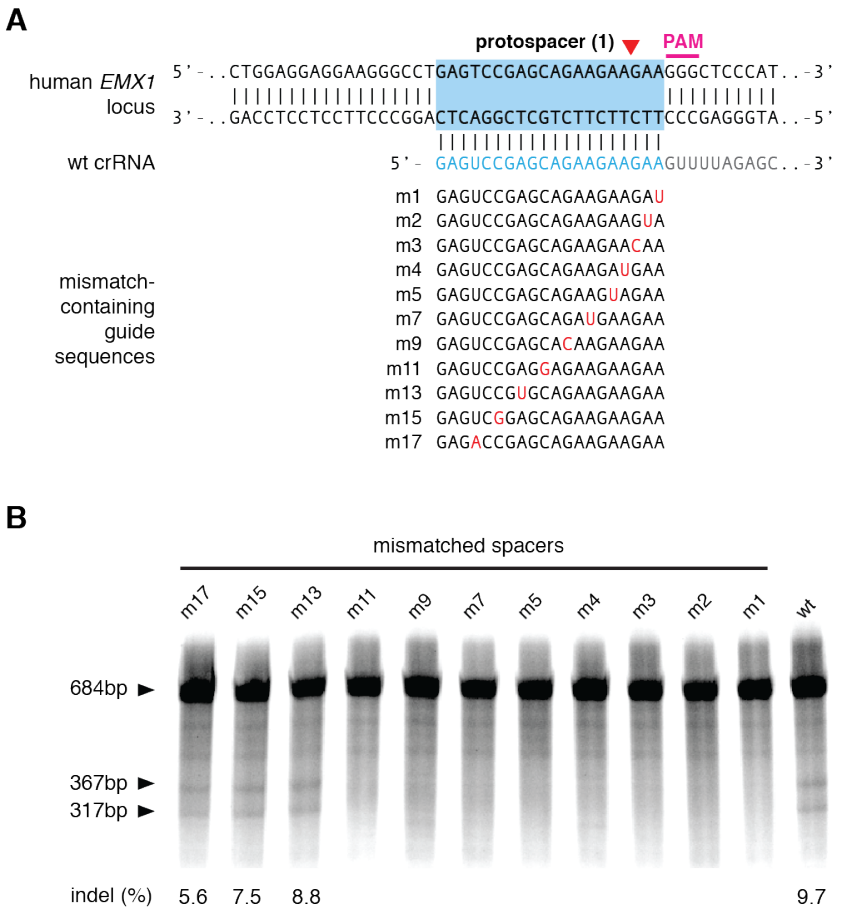


Figure 2-4 Evaluation of Cas9 specificity

(A) *EMX1*-targeting chimeric crRNAs with single point mutations were generated to evaluate the effects of spacer-protospacer mismatches. (B) SURVEYOR assay comparing the cleavage efficiency of different mutant chimeric RNAs.

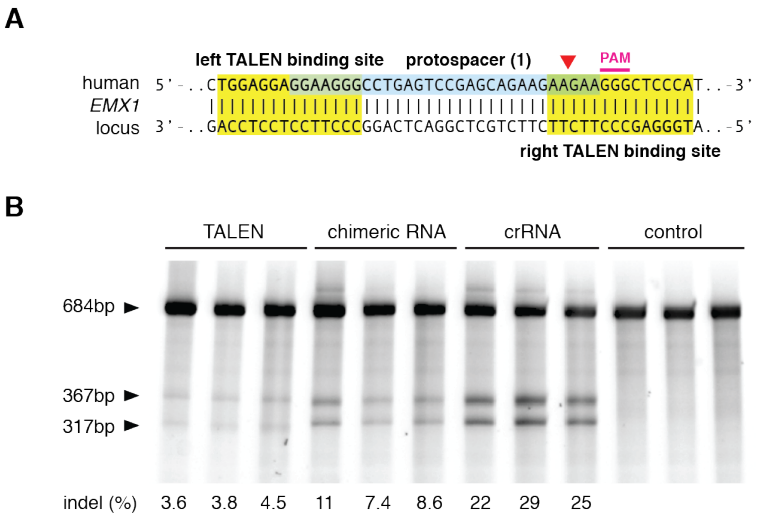


Figure 2-5 Comparison of TALEN and Cas9 efficiency

Schematic showing the design of TALENs targeting *EMX1* and SURVEYOR gel comparing the efficiency of TALEN and SpCas9 ($N = 3$).

Targeted modification of genomes ideally avoids mutations arising from the error-prone NHEJ mechanism. The wild-type SpCas9 is able to mediate site-specific DSBs, which can be repaired through either NHEJ or homology-directed repair (HDR). We engineered an aspartate-to-alanine substitution (D10A) in the RuvC I domain of SpCas9 to convert the nuclease into a DNA nickase (SpCas9n, Figure 2-6A) (*13, 14, 21*), because nicked genomic DNA is typically repaired either seamlessly or through high-fidelity HDR. SURVEYOR (Figure 2-6B) and sequencing of 327 amplicons did not detect any indels induced by SpCas9n. However, it is worth noting that nicked DNA can in rare cases be processed via a DSB intermediate and result in a NHEJ event (*28*). We then tested Cas9-mediated HDR at the same *EMX1* locus with a homology repair template to introduce a pair of restriction sites near the protospacer (Figure 2-6C). SpCas9 and SpCas9n catalyzed integration of the repair template into *EMX1* locus at similar levels (Figure 2-6D), which we further verified via Sanger sequencing (Figure 2-6E). These results demonstrate the utility of CRISPR for facilitating targeted genomic insertions. Given the 14-bp (12-bp from the seed sequence and 2-bp from PAM) target specificity (Figure 2-4B) of the wild type SpCas9, the use of a nickase may reduce off-target mutations.

Finally, the natural architecture of CRISPR loci with arrayed spacers suggests the possibility of multiplexed genome engineering. Using a single CRISPR array encoding a pair of *EMX1*- and *PVALB*-targeting spacers, we detected efficient cleavage at both loci (Figure 2-7A). We further tested targeted deletion of larger genomic regions through concurrent DSBs using spacers against two targets within *EMX1* spaced by 119-bp, and observed a 1.6% deletion efficacy (3 out of 182 amplicons; Figure 2-7B), thus demonstrating the CRISPR system can mediate multiplexed editing within a single genome.

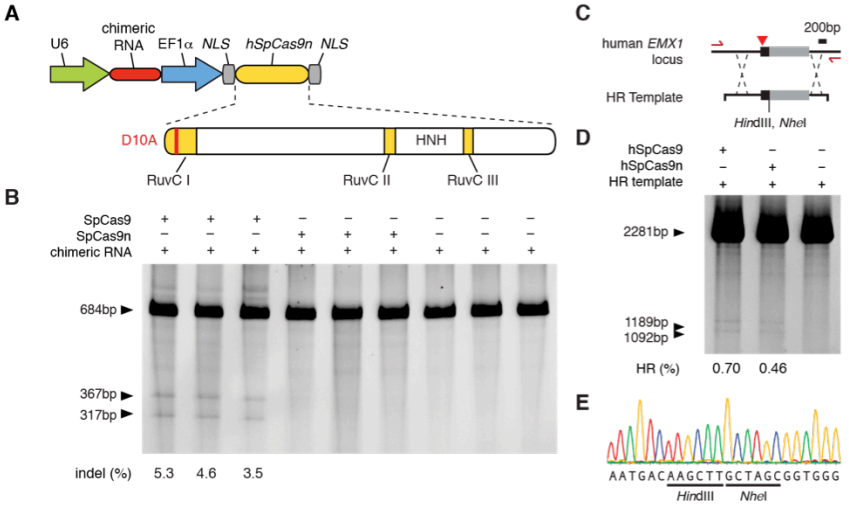


Figure 2-6 SpCas9 applications for homologous recombination

(A) Mutation of the RuvC I domain converts Cas9 into a nicking enzyme (SpCas9n) (B) Co- expression of *EMX1*-targeting chimeric RNA with SpCas9 leads to indels, whereas SpCas9n does not (N = 3). (C) Schematic representation of the recombination strategy. A repair template is designed to insert restriction sites into *EMX1* locus. Primers used to amplify the modified region are shown as red arrows. (D) Restriction fragments length polymorphism gel analysis. Arrows indicate fragments generated by *Hin*dIII digestion. (E) Example chromatogram showing successful recombination.

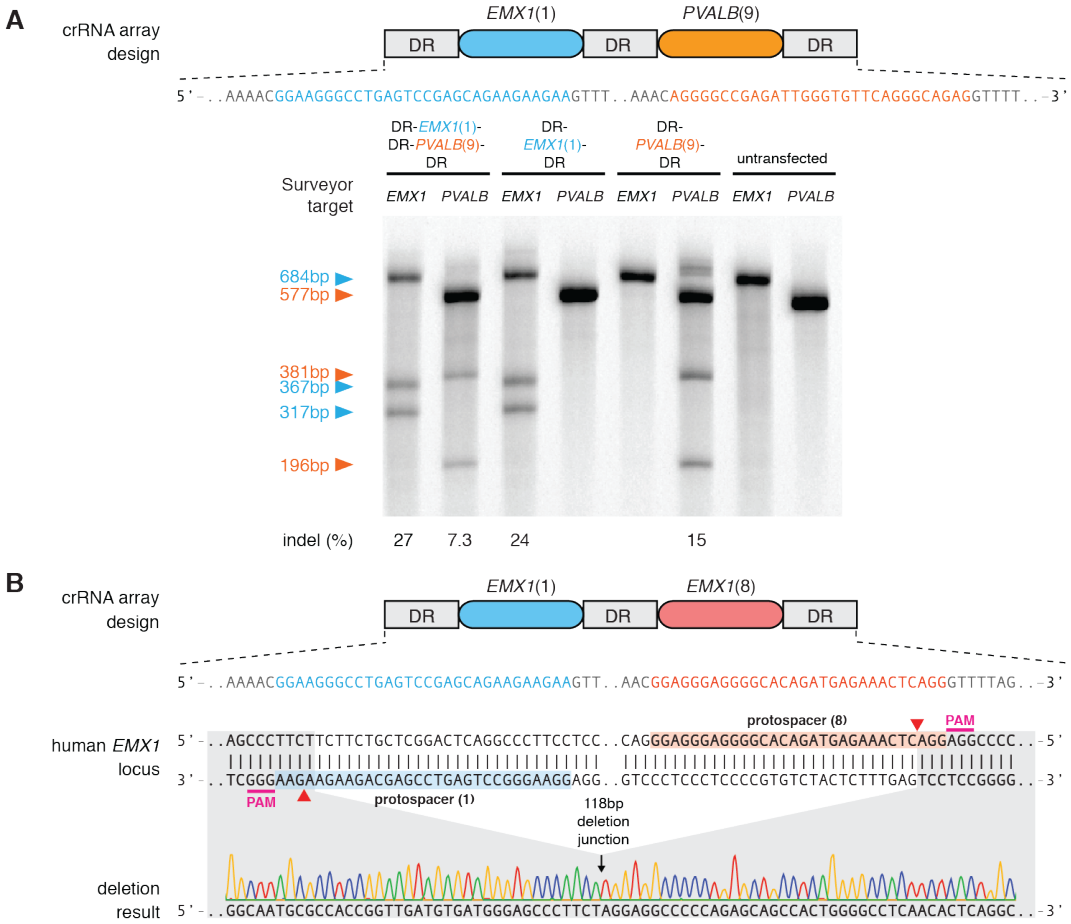


Figure 2-7 SpCas9 mediates multiplexed genome editing

(**A**) SpCas9 can facilitate multiplex genome modification using a crRNA array containing two spacers targeting *EMX1* and *PVALB*. Schematic showing the design of the crRNA array (top). Both spacers mediate efficient protospacer cleavage (bottom). (**B**) SpCas9 can be used to achieve precise genomic deletion. Two spacers targeting *EMX1* (top) mediated a 118bp genomic deletion (bottom).

The ability to use RNA to program sequence-specific DNA cleavage defines a new class of genome engineering tools. Here, we have shown that the *S. pyogenes* CRISPR system can be heterologously reconstituted in mammalian cells to facilitate efficient genome editing; an accompanying study has independently confirmed high efficiency CRISPR-mediated genome targeting in several human cell lines(*29*). However, several aspects of the CRISPR system can be further improved to increase its efficiency and versatility. The requirement for an NGG PAM restricts the *S. pyogenes* CRISPR target space to every 8-bp on average in the human genome (Supplementary Figure 6), not accounting for potential constraints posed by crRNA secondary structure or genomic accessibility due to chromatin and DNA methylation states. Some of these restrictions may be overcome by exploiting the family of Cas9 enzymes and its differing PAM requirements (*23, 24*) across the microbial diversity (*18*). Indeed, other CRISPR loci are likely to be transplantable into mammalian cells; for example, the *Streptococcus thermophilus* LMD-9 CRISPR1 can also mediate mammalian genome cleavage (Supplementary Figure 7). Finally, the ability to carry out multiplex genome editing in mammalian cells enables powerful applications across basic science, biotechnology, and medicine (*30*).

Optimization of sgRNA and characterization of SpCas9 methylation sensitivity

Although a single guide RNA (sgRNA) design consisting of a truncated crRNA and tracrRNA had been previously shown to mediate efficient cleavage *in vitro(13)*, it failed to achieve detectable cleavage at several loci that were efficiently modified by crRNA-tracrRNA duplexes bearing identical guide sequences(*31*). Because the major difference between this sgRNA design and the native crRNA-tracrRNA duplex is the length of the tracrRNA sequence, we tested whether extension of the tracrRNA tail would improve SpCas9 activity.

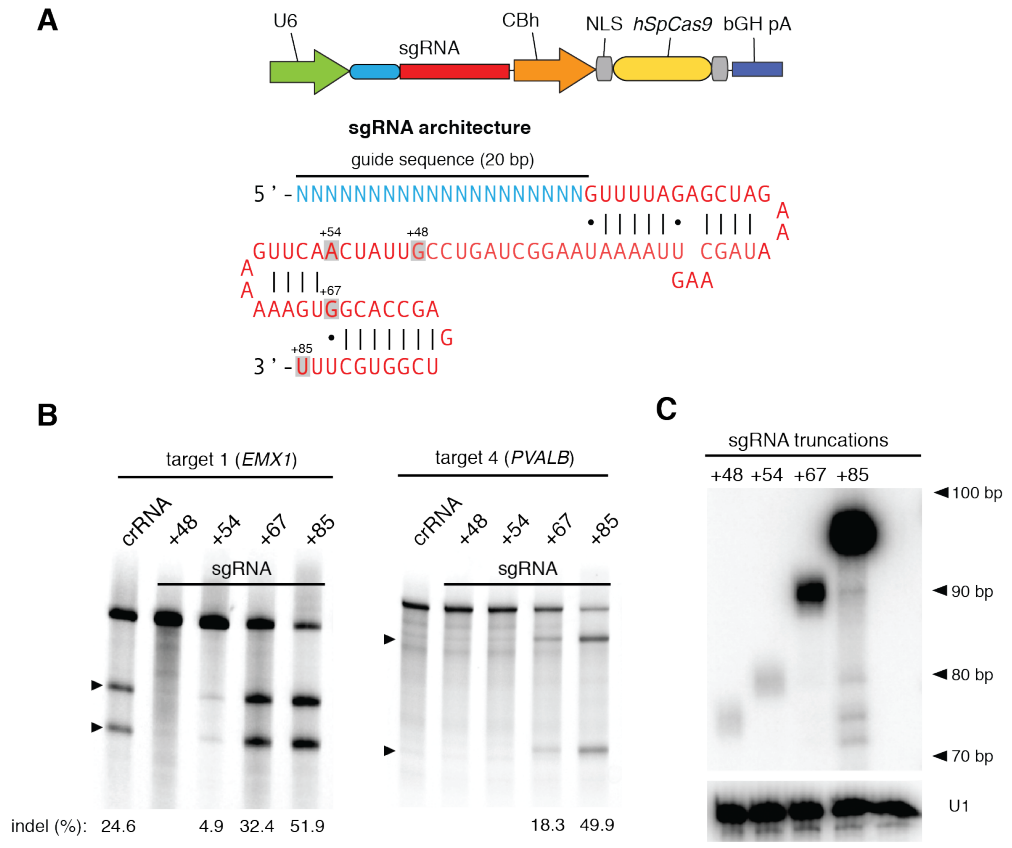


Figure 2-8 Optimization of guide RNA architecture for SpCas9-mediated mammalian genome editing

(A) Schematic of bicistronic expression vector (PX330) for U6 promoter-driven single guide RNA (sgRNA) and CBh promoter-driven human codon-optimized *Streptococcus pyogenes* Cas9 (*hSpCas9*) used for all subsequent experiments. The sgRNA consists of a 20-nt guide sequence (blue) and scaffold (red), truncated at various positions as indicated. (B) SURVEYOR assay for SpCas9-mediated indels at the human *EMX1* and *PVALB* loci. Arrows indicate the expected SURVEYOR fragments (n = 3). (C) Northern blot analysis for the four sgRNA truncation architectures, with U1 as loading control.

We generated a set of sgRNAs targeting multiple sites within the human *EMX1* and *PVALB* loci with different tracrRNA 3' truncations (Figure 2-8A). Using the SURVEYOR nuclease assay(*32*), we assessed the ability of each Cas9-sgRNA complex to generate indels in HEK 293FT cells through the induction of DNA double-stranded breaks (DSBs) and subsequent non-homologous end joining (NHEJ) DNA damage repair. sgRNAs with +67 or +85 nucleotide (nt) tracrRNA tails mediated DNA cleavage at all target sites tested, with up to 5-fold higher levels of indels than the corresponding crRNA-tracrRNA duplexes (Figure 2-8B, Supplementary Figure 9A). Furthermore, both sgRNA designs efficiently modified *PVALB* loci that were previously not targetable using

crRNA-tracrRNA duplexes(*31*) (Figure 2-8B and Supplementary Figure 9B). For all five tested targets, we observed a consistent increase in modification efficiency with increasing tracrRNA length. We performed Northern blots for the guide RNA truncations and found increased levels of expression for the longer tracrRNA sequences, suggesting that improved target cleavage was due to higher sgRNA expression or stability (Figure 2-8C). Furthermore, co-transfection of cells with Cas9 and two optimized sgRNAs with the +85 tail led to robust cleavage of both targets (Supplementary Figure 8A). Using a pair of (+85) sgRNAs flanking an exon in the human *EMX1* gene, we further saw highly efficient deletion of the exon, resulting in approximately ~50% and ~10% of transfected clones with mono- and bi-allelic deletions, respectively (Supplementary Figure 8B, C). Taken together, these data indicate that the tracrRNA tail is important for optimal SpCas9 expression and activity *in vivo.*

To explore whether the genome targeting ability of sgRNA(+85) is influenced by epigenetic factors(*33, 34*) that constrain the alternative transcription activator-like effector nuclease (TALENs)(*5, 35-38*) and potentially also zinc finger nuclease (ZFNs)(*1-4, 39*) technologies, we further tested the ability of SpCas9 to cleave methylated DNA. Using either unmethylated or M.SssI-methylated pUC19 as DNA targets (Figure 2-9A, B) in a cell-free cleavage assay, we showed that SpCas9 efficiently cleaves pUC19 regardless of CpG methylation status in either the 20-bp target sequence or the PAM (Figure 2-9C). To test whether this is also true in vivo, we designed sgRNAs to target a highly methylated region of the human SERPINB5 locus (Figure 2-10A, B). All three sgRNAs tested were able to mediate indel mutations in endogenously methylated targets (Figure 2-10C).

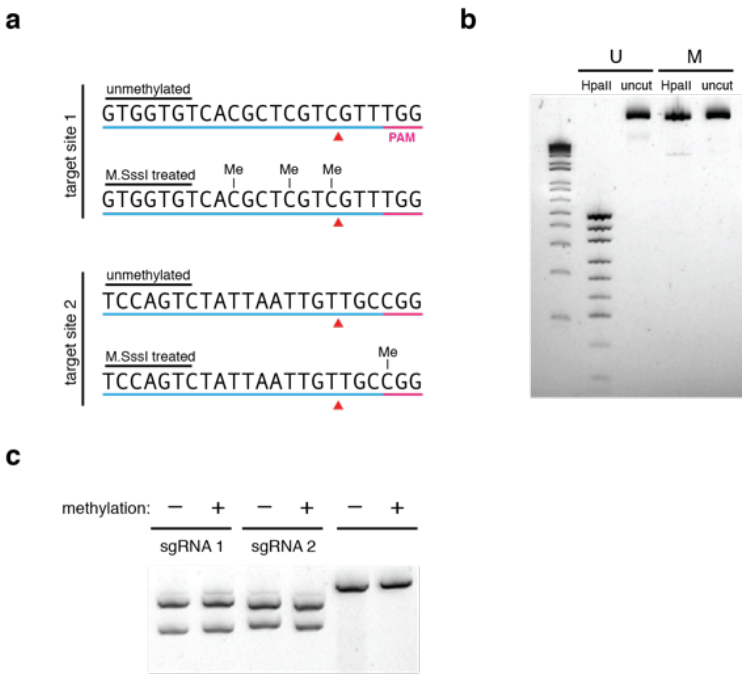


Figure 2-9 Cas9 targeting of methylated DNA in vitro

(A) Sequence of CpG dinucleotide-containing targets in pUC19 plasmid methylated *in vitro* by M.SssI. Methyl-CpGs in either the target sequence or PAM are indicated; arrows indicate expected cleavage site. (B) Unmethylated (U) or methylated (M) pUC19 was subjected to restriction digest by the methylation-sensitive restriction enzyme HpaII. Unmethylated pUC19 is digested into a ladder while M.SssI-treated pUC19 is protected from HpaII digestion. (C) Cleavage of either unmethylated or methylated targets 1 and 2 on linearized pUC19 by SpCas9. No sgRNAs are present in negative control lanes.

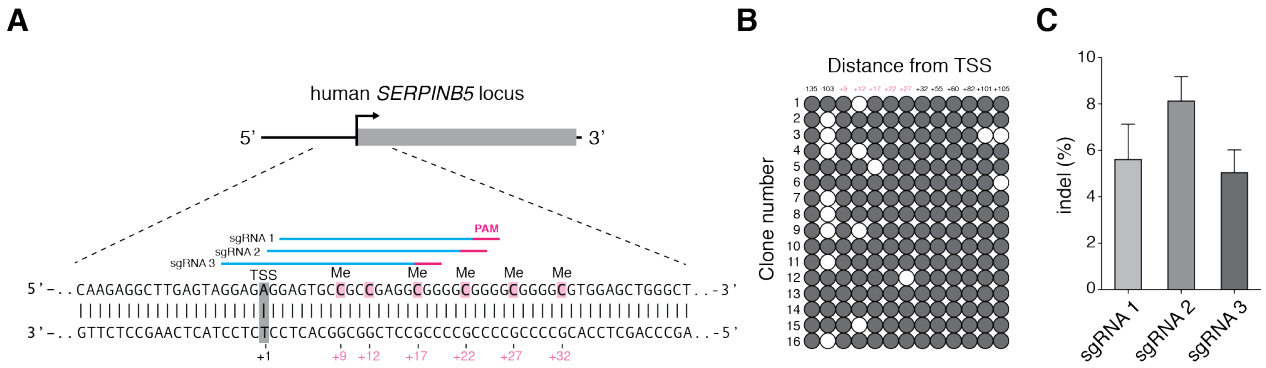


Figure 2-10 Cas9 targeting of methylated DNA in vivo

(A) Schematic of the human *SERPINB5* locus. sgRNAs and PAMs are indicated by colored bars above sequence; methylcytosine (Me) are highlighted (pink) and numbered relative to the transcriptional start site (TSS, +1). (B) Methylation status of *SERPINB5* assayed by bisulfite sequencing of 16 clones. Filled circles, methylated CpG; open circles, unmethylated CpG. (C) Modification efficiency by three sgRNAs targeting the methylated region of *SERPINB5,* assayed by deep sequencing ($n = 2$). Error bars indicate Wilson intervals.

Single-stranded DNA repair templates for high efficiency gene modification

Having optimized the sgRNA scaffold for Cas9 cleavage, we sought to further improve Cas9-mediated HDR. Traditionally, targeted DNA modifications have required use of plasmid-based donor repair templates that contain homology arms flanking the site of alteration(*40, 41*) (Figure 1-1). The homology arms on each side can vary in length, but are typically longer than 500-bp(*41, 42*). This method can be used to generate large modifications, including insertion of reporter genes such as fluorescent proteins or antibiotic resistance markers. The design and construction of targeting plasmids has been described elsewhere(*43*).

More recently, single-stranded DNA oligonucleotides (ssODNs) have been used in place of targeting plasmids for short modifications within a defined locus without cloning(*44*). To achieve high HDR efficiencies, ssODNs contain flanking sequences of at least 40 bp on each side that are homologous to the target region, and can be oriented in either the sense or antisense direction relative to the target locus.

We sought to test here the use of ssODNs and targeting vector to mediate HDR (Figure 2-11, B) with both wildtype and nickase mutant of Cas9 in HEK 293FT and HUES9 cells (Figure 2-11C). We have not been able to detect HDR in HUES9 cells using the Cas9 nickase, which may be due to low efficiency or a potential difference in repair activities in HUES9 cells. It is worth noting that there is some variability in the cleavage efficiency of a given sgRNA, and on rare occasions certain sgRNAs may not work for reasons yet unknown.

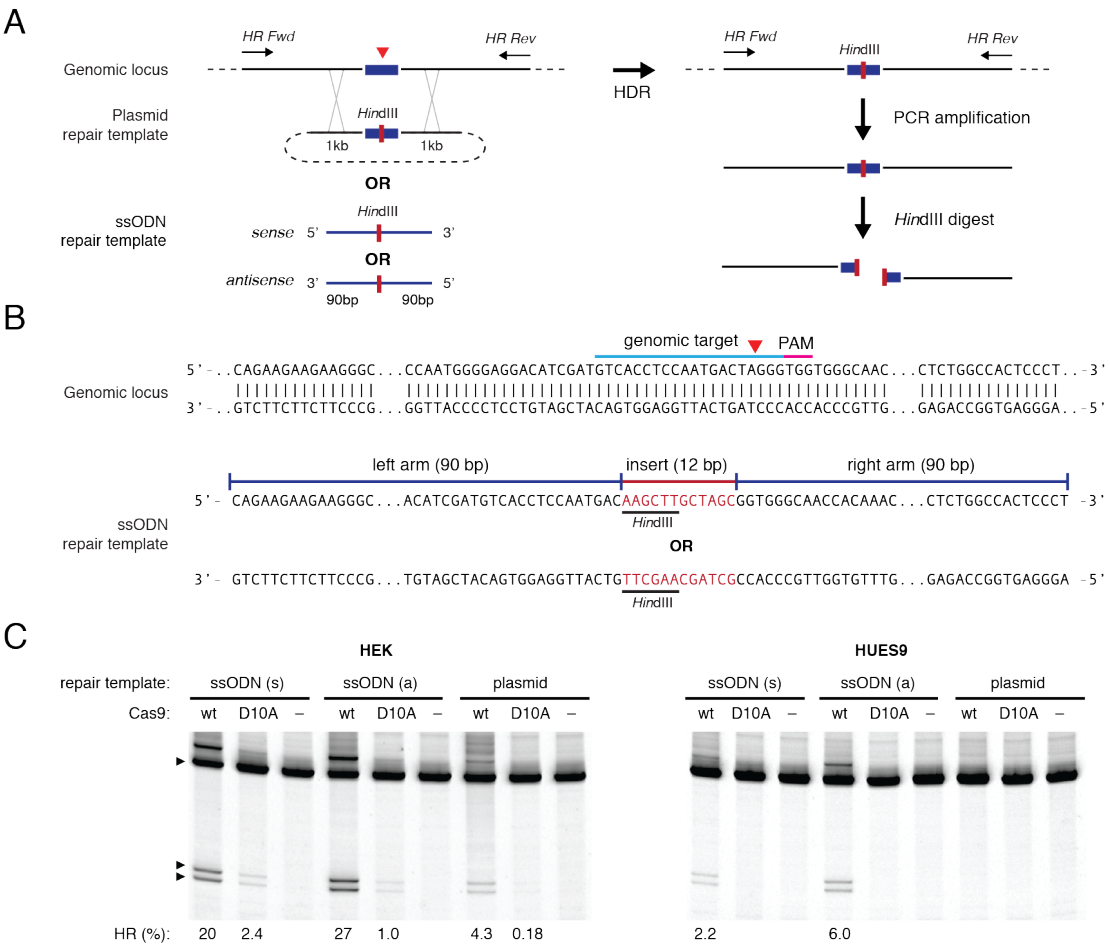


Figure 2-11 Single-stranded oligonucleotide (ssODN)-mediated HDR

(A) A targeting plasmid or ssODNs (sense or antisense) with homology arms can be used to edit the sequence (red bar) at a target genomic locus cleaved by Cas9 (red triangle). To assay the efficiency of HDR, a *Hin*dIII site was introduced into the target locus and the target locus was PCR amplified using primers that anneal outside of the region of homology. Digestion of the PCR amplicon with *Hin*dIII reveals the occurrence of HDR events. (B) ssODNs, oriented in either the sense or the antisense direction relative to locus of interest, can be used in combination with Cas9 to achieve efficient HDR-mediated editing at the target locus. A minimal homology region of 40-bp, preferably 90-bp, is recommended on either side of modification (red bar). (C) Example ssODNs for recombination in the *EMX1* locus are shown. Each ssODN contains homology arms of 90-bp each flanking a 12-bp insertion of two restriction sites.

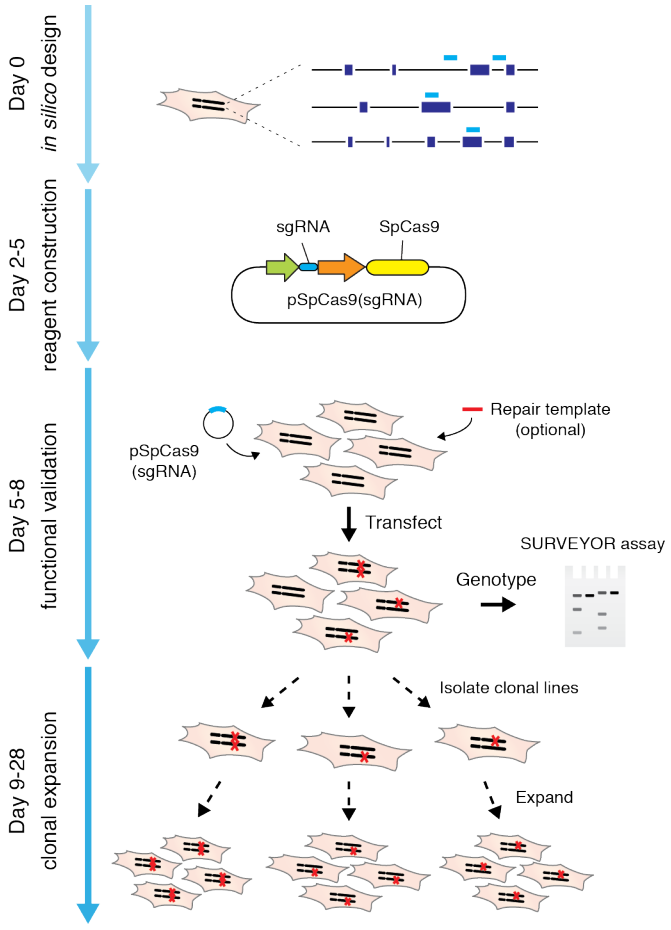


Figure 2-12 Pipeline for rapid generation of cell lines using Cas9

Steps for reagent design, construction, validation, and cell line expansion. Custom sgRNAs (light blue bars) for each target, as well as genotyping primers, are designed *in silico*. sgRNA expression vectors are then cloned into a plasmid containing Cas9 (PX330) and verified via DNA sequencing. Completed plasmids (pCRISPRs), and optional repair templates for facilitating homology directed repair, are then transfected into cells and assayed for ability to mediate targeted cleavage. Finally, transfected cells can be clonally expanded to derive isogenic cell lines with defined mutations.

Discussion

Here, we have presented a human codon-optimized, nuclear localization sequence-flanked Cas9 to facilitate gene editing in mammalian cells. We have demonstrated multiplexed cleavage of endogenous loci using Cas9 guided by both crRNA:tracrRNA duplex and a chimeric single guide RNA. We've further engineered a mutant Cas9 carrying a disruption in one of its catalytic domains(*13, 14*) to nick rather than cleave DNA, allowing for single-stranded breaks and preferential repair through HDR(*31*), which could potentially ameliorate unwanted indel mutations from off-target DSBs. Lastly, we have optimized the sgRNA for robust activity across a number of loci previously unable to be targeted, and shown that single-stranded DNA donor templates can be used to mediate HDR at high efficiencies. In summary, we have optimized a pipeline from *in silico* design of target sites to rapidly generating cell lines in the span of a few weeks (Figure 2-12).

Given its ease of implementation and multiplex capability, Cas9 has already been used to generate engineered eukaryotic cells carrying specific mutations via both NHEJ and HDR(*29, 31, 45-47*). In addition, direct injection of sgRNA and mRNA encoding Cas9 into embryos has enabled the rapid generation of transgenic mice with multiple modified alleles(*48, 49*); these results hold immense promise for editing organisms that are otherwise genetically intractable.

References

1. M. H. Porteus, D. Baltimore, Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (May 2, 2003).

2. J. Miller *et al.*, An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**, 778 (2007).

3. J. Sander *et al.*, Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nature methods* **8**, 67 (2011).

4. A. J. Wood *et al.*, Targeted genome editing across species using ZFNs and TALENs. *Science* **333**, 307 (Jul 15, 2011).

5. M. Christian *et al.*, Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757 (Oct, 2010).

6. F. Zhang *et al.*, Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **29**, 149 (2011).

7. J. Miller *et al.*, A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29**, 143 (2011).

8. D. Reyon *et al.*, FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology* **30**, 460 (2012).

9. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)* **326**, 1509 (2009).

10. M. Moscou, A. Bogdanove, A simple cipher governs DNA recognition by TAL effectors. *Science (New York, N.Y.)* **326**, 1501 (2009).

11. A. Hinnen, J. Hicks, G. Fink, Transformation of yeast. *Proceedings of the National Academy of Sciences of the United States of America* **75**, 1929 (1978).

12. B. L. Stoddard, Homing endonuclease structure and function. *Quarterly reviews of biophysics* **38**, 49 (Feb, 2005).

13. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816 (2012).

14. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579 (Sep 25, 2012).

15. J. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67 (2010).

16. H. Deveau, J. Garneau, S. Moineau, CRISPR/Cas system and its role in phage-bacteria interactions. *Annual review of microbiology* **64**, 475 (2010).

17. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167 (Jan 8, 2010).

18. K. Makarova *et al.*, Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology* **9**, 467 (2011).

19. D. Bhaya, M. Davison, R. Barrangou, CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics* **45**, 273 (2011).

20. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602 (2011).

21. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39**, 9275 (2011).

22. A. H. Magadan, M. E. Dupuis, M. Villion, S. Moineau, Cleavage of phage DNA by the Streptococcus thermophilus CRISPR3-Cas system. *PloS one* **7**, e40913 (2012).

23. H. Deveau *et al.*, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of bacteriology* **190**, 1390 (2008).

24. F. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, England)* **155**, 733 (2009).

25. M. Jinek, J. A. Doudna, A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457**, 405 (Jan 22, 2009).

26. C. D. Malone, G. J. Hannon, Small RNAs as guardians of the genome. *Cell* **136**, 656 (Feb 20, 2009).

27. G. Meister, T. Tuschl, Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**, 343 (Sep 16, 2004).

28. M. T. Certo *et al.*, Tracking genome engineering outcome at individual DNA breakpoints. *Nature methods* **8**, 671 (Aug, 2011).

29. P. Mali *et al.*, RNA-guided human genome engineering via Cas9. *Science* **339**, 823 (Feb 15, 2013).

30. P. A. Carr, G. M. Church, Genome engineering. *Nature biotechnology* **27**, 1151 (Dec, 2009).

31. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

32. D. Y. Guschin *et al.*, A rapid and general assay for monitoring endogenous gene modification. *Methods in molecular biology* **649**, 247 (2010).

33. S. Bultmann *et al.*, Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. *Nucleic acids research* **40**, 5368 (Jul, 2012).

34. J. Valton *et al.*, Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *J Biol Chem* **287**, 38427 (Nov 9, 2012).

35. D. Hockemeyer *et al.*, Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**, 731 (Aug, 2011).

36. C. Mussolino, R. Morbitzer, F. Lütge ... , A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids ...* , (2011).

37. P. D. Hsu, F. Zhang, Dissecting neural function using targeted genome engineering technologies. *ACS chemical neuroscience* **3**, 603 (Aug 15, 2012).

38. N. E. Sanjana *et al.*, A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* **7**, 171 (Jan, 2012).

39. S. Bobis-Wozowicz, A. Osiak, S. H. Rahman, T. Cathomen, Targeted genome editing in pluripotent stem cells using zinc-finger nucleases. *Methods* **53**, 339 (Apr, 2011).

40. O. Smithies, R. G. Gregg, S. S. Boggs, M. A. Koralewski, R. S. Kucherlapati, Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* **317**, 230 (Sep 19-25, 1985).

41. K. Thomas, K. Folger, M. Capecchi, High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419 (1986).

42. P. Hasty, J. Rivera-Perez, A. Bradley, The length of homology required for gene targeting in embryonic stem cells. *Mol Cell Biol* **11**, 5586 (Nov, 1991).

43. S. Wu, G. Ying, Q. Wu, M. R. Capecchi, A protocol for constructing gene targeting vectors: generating knockout mice for the cadherin family and beyond. *Nat Protoc* **3**, 1056 (2008).

44. F. Chen *et al.*, High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nature methods* **8**, 753 (2011).

45. M. Jinek *et al.*, RNA-programmed genome editing in human cells. *eLife* **2**, (2013).

46. S. W. Cho, S. Kim, J. M. Kim, J. S. Kim, Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology* **31**, 230 (Mar, 2013).

47. W. Hwang *et al.*, Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology* **31**, 227 (2013).

48. H. Wang *et al.*, One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910 (May 9, 2013).

49. B. Shen *et al.*, Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. *Cell research* **23**, 720 (May, 2013).

Chapter 3 Double Nicking by Cas9 for enhanced genome editing specificity

F. Ann Ran[1,2,*], Patrick D. Hsu[1,2,*] Chie-Yu Lin[1,3], Jonathan S. Gootenberg[1], Silvana Konermann[1], Alexandro E. Trevino[1], David A. Scott[1], Azusa Inoue[4], Shogo Matoba[4], Yi Zhang[4], and Feng Zhang[1]

[1] Broad Institute and McGovern Institutes,
Department of Brain and Cognitive Sciences,
and Department of Biological Engineering
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] Department of Molecular and Cellular Biology,
Harvard University, Cambridge, MA 02138, USA

[3] Harvard/MIT Division of Health Sciences and Technology
[4] Howard Hughes Medical Institute,
Program in Cellular and Molecular Medicine,
and Department of Genetics
Harvard Stem Cell Institute
Harvard Medical School, Boston, MA 02115, USA

*These authors contributed equally to this work

Acknowledgements

We thank Joshua Weinstein and Yinqing Li for statistical consultation, Xuebing Wu and Phillip Sharp for assistance with northern blotting experiments. This work was supported by a NIH Director's Pioneer Award (DP1-MH100706) to F.Z., the Keck, McKnight, Gates, Damon Runyon, Searle Scholars, Klingenstein, and Simons Foundations, Bob Metcalfe, Mike Boylan, and Jane Pauley.

Author contribution

F.A.R., P.D.H., and F.Z. conceived the study, F.A.R., P.D.H., C.Y.L., and F.Z. designed and performed the experiments with assistance from all authors. F.Z. supervised all aspects of the study. F.A.R. P.D.H., C.Y.L., A.E.T., and F.Z. wrote the manuscripts with help from all authors. J.S.G. and D.A.S. contributed computational analysis for deep sequencing. A.I., S.M., and Y.Z. contributed blastocyst injections. In particular, I designed and conducted the sgRNA extension experiments, double nicking and homologous recombination experiments, carried out off-target analysis.

This chapter contains work from the manuscript "Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity," published in *Cell*, September 12, 2013, Vol. 154 pp. 1380-1389. The text and figures were modified to fit the format of this dissertation.

Introduction

The ability to perturb the genome in a precise and targeted fashion is crucial for understanding genetic contributions to biology and disease. Genome engineering of cell lines or animal models has traditionally been accomplished through random mutagenesis or low-efficiency gene targeting. To facilitate genome editing, programmable sequence-specific DNA nuclease technologies have enabled targeted modification of endogenous genomic sequences with high efficiency, particularly in species that have proven genetically intractable (*1-4*). The RNA-guided Cas9 nucleases from the microbial CRISPR (clustered regularly interspaced short palindromic repeat)-Cas systems are robust and versatile tools for stimulating targeted double-stranded DNA breaks (DSBs) in eukaryotic cells (*5-15*), where the resulting cellular repair mechanisms – non-homologous end joining (NHEJ) or homology-directed repair (HDR) pathways – can be exploited to induce error-prone or defined alterations (*16-18*).

The Cas9 nuclease from *Streptococcus pyogenes* can be directed by a chimeric single guide RNA (sgRNA) (*19*) to any genomic locus preceding a 5'-NGG protospacer-adjacent motif (PAM). A 20-nt guide sequence within the sgRNA directs Cas9 to the genomic target via Watson-Crick base pairing and can be easily programmed to target a desired genomic locus (*9, 10, 19, 20*). Recent studies of Cas9 specificity have demonstrated that although each base within the 20-nt guide sequence contributes to overall specificity, multiple mismatches between the guide RNA and its complementary target DNA sequence can be tolerated depending on the quantity, position, and base identity of mismatches (*5, 21-23*), leading to potential off-target DSBs and indel formation. These unwanted mutations can potentially limit the utility of Cas9 for genome editing applications that require high levels of precision, such as generation of isogenic cell lines for testing causal genetic variations (*24*) or *in vivo* and *ex vivo* genome editing-based therapies.

To improve the specificity of Cas9-mediated genome editing, we developed a novel strategy that combines the D10A mutant nickase version of Cas9 (Cas9n) (*5, 19, 20*) with a pair of offset sgRNAs complementary to

opposite strands of the target site. While nicking of both DNA strands by a pair of Cas9 nickases leads to site-specific DSBs and NHEJ, individual nicks are predominantly repaired by the high-fidelity base excision repair pathway (BER) (*25*). A paired nickase strategy was described while this manuscript was under review which suggests the possibility for engineering a system to ameliorate off-target activity (*26*). In a manner analogous to dimeric zinc finger nucleases (ZFNs) (*27-30*) and transcription activator-like effector nucleases (TALENs) (*30-36*), where DNA cleavage requires synergistic interaction of two independent specificity-encoding DNA-binding modules directing FokI nuclease monomers, this double nicking strategy minimizes off-target mutagenesis by each individual Cas9n-sgRNA complex while maintaining on-target modification rates similar to those of wild type Cas9. Here we define crucial parameters for the selection of sgRNA pairs that facilitate effective double nicking, compare the specificity of wildtype Cas9 and Cas9n with double nicking, and demonstrate a variety of experimental applications that can be achieved using double nicking in cells as well as in mouse zygotes.

Extension of guide sequence does not improve Cas9 targeting specificity

Cas9 targeting is facilitated by base-pairing between the 20-nt guide sequence within the sgRNA and the target DNA (*9, 10, 19, 20*). We reasoned that cleavage specificity might be improved by increasing the length of base-pairing between the guide RNA and its target locus. To test this, we generated U6-driven expression cassettes (*22*) to express three sgRNAs with 20-nt (sgRNA 1) or 30-nt guide sequences (sgRNAs 2 and 3) targeting a locus within the human *EMX1* gene (Figure 3-1A).

We and others have previously shown that while single-base mismatches between the PAM-distal region of the guide sequence and target DNA are well-tolerated by Cas9, multiple mismatches in this region can significantly affect on-target activity (*21, 22, 37, 38*). To determine whether additional PAM-distal bases (21-30) could influence overall targeting specificity, we designed sgRNAs 2 and 3 to contain 10 additional bases consisting of

either 10 perfectly matched or 8 mismatched bases (bases 21-28). Surprisingly, we observed that these extended sgRNAs mediated similar levels of modification at the target locus in HEK 293FT cells regardless of whether the additional bases were complementary to the genomic target (Figure 3-1B). Subsequent Northern blots revealed that the majority of both sgRNA 2 and 3 were processed to the same length as sgRNA 1, which contains the same 20-nt guide sequence without additional bases (Figure 3-1C).

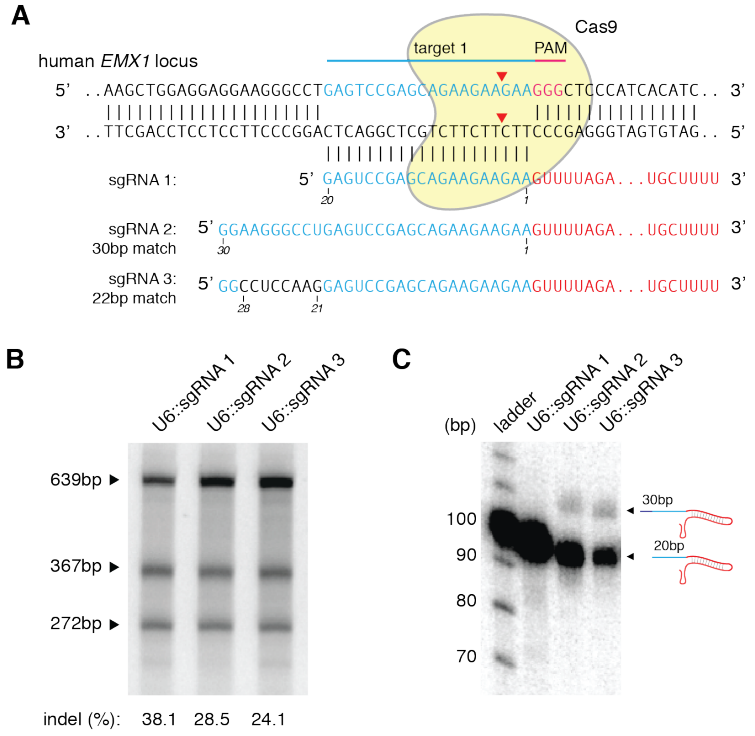


Figure 3-1 Effect of guide sequence extension on Cas9 activity

(A) Schematic showing Cas9 with matching or mismatching sgRNA sequences targeting a locus (target 1) within the human *EMX1* gene. (B) SURVEYOR assay gel showing comparable modification of target 1 by sgRNAs bearing 20- and 30-nt long guide sequences. (C) Northern blot showing that extended sgRNAs are largely reverted to 20-nt guide-length sgRNAs in HEK 293FT cells.

Cas9 nickase generates efficient NHEJ with paired, offset guide RNAs

Given that extension of the guide sequence failed to improve Cas9 targeting specificity, we sought an alternative strategy for increasing the overall base-pairing length between the guide sequence and its DNA target. Cas9 enzymes contain two conserved nuclease domains, HNH and RuvC, which cleave the DNA strand complementary and non-complementary to the guide RNA, respectively. Mutations of the catalytic residues (D10A in RuvC and H840A in HNH) convert Cas9 into DNA nickases (*5, 19, 20*). As single-strand nicks are preferentially repaired by the high-fidelity BER pathway (*25*), we reasoned that two Cas9 nicking enzymes directed by a pair of sgRNAs targeting opposite strands of a target locus could mediate DSBs while minimizing off-target activity (Figure 3-2A).

A number of factors may affect cooperative nicking leading to indel formation, including steric hindrance between two adjacent Cas9 molecules or Cas9-sgRNA complexes, overhang type, and sequence context; some of these may be characterized by testing multiple sgRNA pairs with distinct target sequences and offsets (the distance between the PAM-distal (5') ends of the guide sequence of a given sgRNA pair). To systematically assess how sgRNA offsets might affect subsequent repair and generation of indels, we first designed sets of sgRNA pairs targeted against the human *EMX1* genomic locus separated by a range of offset distances from approximately -200 to 200 bp to create both 5'- and 3'-overhang products (Figure 3-2A, Supplementary Table 2). We then assessed the ability of each sgRNA pair, with the D10A Cas9 mutant (referred to as Cas9n; H840A Cas9 mutant is referred to as Cas9H840A), to generate indels in human HEK 293FT cells. Robust NHEJ (up to 40%) was observed for sgRNA pairs with offsets from -4 to 20 bp, with modest indels forming in pairs offset by up to 100-bp (Figure 3-2B, left panel). We subsequently recapitulated these findings by testing similarly offset sgRNA pairs at two other genomic loci, *DYRK1A* and *GRIN2B* (Figure 3-2B, right panels). Notably, across all three loci examined, only

sgRNA pairs creating 5' overhangs with less than 8bp overlap between the guide sequences (offset greater than -8 bp) were able to mediate detectable indel formation (Figure 3-2C). Importantly, each guide used in these assays is able to efficiently induce indels when paired with wildtype Cas9 (Supplementary Table 2), indicating that the relative positions of the guide pairs are the most important parameters in predicting double nicking activity.

Since Cas9n and Cas9H840A nick opposite strands of DNA, substitution of Cas9n with Cas9H840A with a given sgRNA pair should result in the inversion of the overhang type. For example, a pair of sgRNAs that will generate a 5' overhang with Cas9n should in principle generate the corresponding 3' overhang instead. Therefore, sgRNA pairs that lead to the generation of a 3' overhang with Cas9n might be used with Cas9H840A to generate a 5' overhang. Unexpectedly, we tested Cas9H840A with a set of sgRNA pairs designed to generate both 5' and 3' overhangs (offset range from –278 to +58 bp), but were unable to observe indel formation. Further work will be needed to identify the necessary design rules for sgRNA pairing to allow double nicking by Cas9H840A.

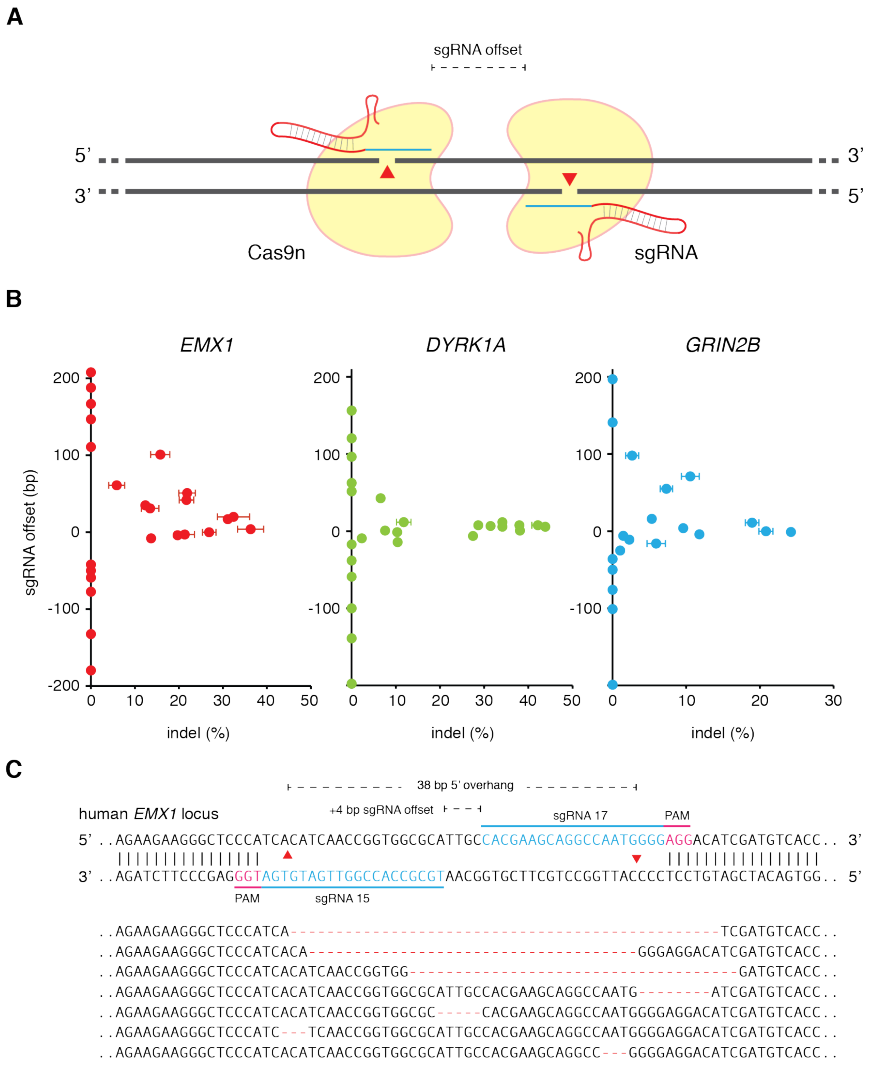


Figure 3-2 Double nicking facilitates efficient genome editing in human cells

(A) Schematic illustrating DNA double-stranded breaks using a pair of sgRNAs guiding Cas9 D10A nickases (Cas9n). The D10A mutation renders Cas9 able to cleave only the strand complementary to the sgRNA; a pair of sgRNA-Cas9n complexes can nick both strands simultaneously. sgRNA offset is defined as the distance between the PAM-distal (5') ends of the guide sequence of a given sgRNA pair; positive offset requires the sgRNA complementary to the top strand (sgRNA a) to be 5' of the sgRNA complementary to the bottom strand (sgRNA b). (B) Efficiency of double nicking induced NHEJ as a function of the offset distance between two sgRNAs. (n = 3; error bars show mean ± s.e.m.) (C) Representative sequences of the human *EMX1* locus targeted by Cas9n. sgRNA target sites and PAMs are indicated by blue and magenta bars respectively. Below, selected sequences showing representative indels.

Double nicking mediates genome editing with improved specificity

Having established that double nicking (DN) mediates high efficiency NHEJ at levels comparable to those induced by wildtype Cas9, we next studied whether DN has improved specificity over wildtype Cas9 by measuring their off-target activities. We co-delivered Cas9n with sgRNAs 1 and 9, spaced by a +23 bp offset, to target the human EMX1 locus in HEK 293FT cells (Figure 3-3A). This DN configuration generated on-target indel levels similar to those generated by the wildtype Cas9 paired with each sgRNA alone (Figure 3-3B, left panel). Strikingly, unlike with wildtype Cas9, DN did not generate detectable modification at a previously validated sgRNA 1 off-target site, OT-4, by SURVEYOR assay (*22*) (Figure 3-3B, right panel), suggesting that DN can potentially reduce the likelihood of off-target modifications. Using deep sequencing to assess modification at 5 different sgRNA 1 off-target loci (Figure 3-3A), we observed significant mutagenesis at all sites with wild type Cas9 + sgRNA 1 (Figure 3-3C). In contrast, cleavage by Cas9n at 5 off-target sites tested was barely detectable above background sequencing error.

Using the ratio of on- to off-target modification levels as a metric of specificity, we found that Cas9n with a pair of sgRNAs was able to achieve over 100-fold greater specificity relative to wild type Cas9 with one of the sgRNAs (Figure 3-3D). We conducted additional off-target analysis by deep sequencing for two sgRNA pairs (offsets of +16 and +20 bp) targeting the *VEGFA* locus, with similar results (Figure 3-3E). DN at these off-target loci was able to achieve 200 to over 1500-fold greater specificity than the wild-type Cas9 (Figure 3-3F, Supplementary Table 2). Taken together, these results demonstrate that Cas9-mediated double nicking minimizes off-target mutagenesis and is suitable for genome editing with increased specificity.

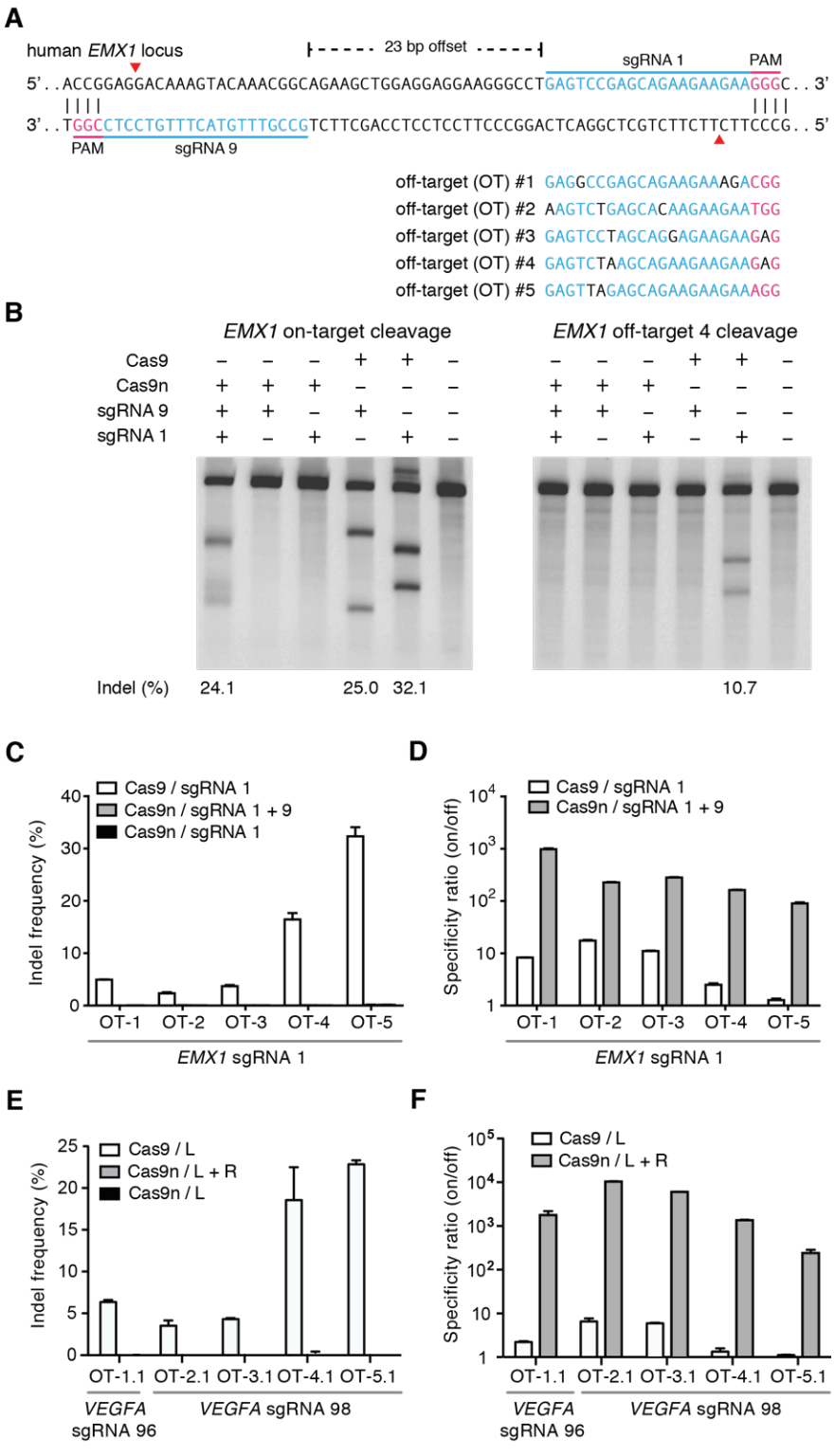


Figure 3-3 Double nicking enables high precision genome editing in human cells

(A) Schematic illustrating Cas9n double nicking (red arrows) the human *EMX1* locus. Five off-target loci with sequence homology to *EMX1* target 1 were selected to screen for Cas9n specificity. (B) On-target modification rate by Cas9n and a pair of sgRNAs is comparable to those mediated by wildtype Cas9 and single sgRNAs (left panel). Cas9-sgRNA 1 complexes generate significant off-target mutagenesis, while no off-target locus modification is detected with Cas9n (right panel). (C) Five off-target loci of sgRNA 1 are examined for indel modifications by deep sequencing of transfected HEK 293FT cells. (n = 3, error bars show mean ± s.e.m.) (D) Specificity comparison of Cas9n with double nicking and wildtype Cas9 with sgRNA 1 alone at the off-target sites. Specificity ratio is calculated as on-target/off-target modification rates. (n = 3; error bars show mean ± s.e.m.) (E, F) Double nicking minimizes off-target modification at two additional human *VEGFA* loci while maintaining high specificity (on/off target modification ratio; n = 3, error bars show mean ± s.e.m.).

Double nicking facilitates high-efficiency homology directed repair, NHEJ-mediated DNA insertion, and genomic microdeletions

DSBs can stimulate homology directed repair (HDR) to enable highly precise editing of genomic target sites. To evaluate DN-induced HDR, we targeted the human *EMX1* locus with pairs of sgRNAs offset by -3 and +18 bp (generating 31- and 52-bp 5' overhangs), respectively, and introduced a single-stranded oligodeoxynucleotide (ssODN) bearing a *Hin*dIII restriction site as the HDR repair template (Figure 3-4A). Each DN sgRNA pair successfully induced HDR at frequencies higher than those of single-guide Cas9n nickases and comparable to those of wild-type Cas9 (Figure 3-4B). Furthermore, genome editing in embryonic stem cells or patient derived induced pluripotent stem cells represents a key opportunity for generating and studying new disease paradigms as well as developing new therapeutics. Since single nick approaches to inducing HDR in human embryonic stem cells (hESCs) have met with limited success (*22*), we attempted DN in the HUES62 hES cell line and observed successful HDR (Figure 3-4C).

To further characterize how offset sgRNA spacing affects the efficiency of HDR, we next tested in HEK 293FT cells a set of sgRNA pairs where the cleavage site of at least one sgRNA is situated near the site of recombination (overlapping with the HDR ssODN donor template arm). We observed that sgRNA pairs generating 5' overhangs and having at least one nick occurring within 22 bp of the homology arm are able to induce HDR at levels

comparable to those of wildtype Cas9-mediated HDR, and significantly greater than those of single Cas9n-sgRNA nicking. In contrast, we did not observe HDR with sgRNA pairs that generated 3'-overhangs or double nicking of the same DNA strand (Figure 3-4D).

The ability to create defined overhangs could enable precise insertion of donor repair templates containing compatible overhangs via NHEJ-mediated ligation (*39*). To explore this alternative strategy for transgene insertion, we targeted the *EMX1* locus with Cas9n and an sgRNA pair designed to generate a 43 bp 5'-overhang near the stop codon, and supplied a double-stranded oligonucleotide (dsODN) duplex with matching overhangs (Figure 3-5A). The annealed dsODN insert, containing multiple epitope tags and a restriction site, was successfully integrated into the target at a frequency of 3% (1/37 screened by Sanger sequencing of cloned amplicons). This ligation-based strategy thus illustrates an effective approach for inserting dsODNs encoding short modifications such as protein tags or recombination sites into an endogenous locus.

Additionally, we targeted combinations of sgRNA pairs (4 sgRNAs per combination) to the DYRK1A locus in HEK 293FT cells to facilitate genomic microdeletions. We generated a set of sgRNAs to mediate 0.5 kb, 1 kb, 2 kb, and 6 kb deletions (Figure 3-5B, Supplementary Table 3: sgRNAs 32, 33, 54-61) and verified successful multiplex nicking-mediated deletion over these ranges via PCR screen of predicted deletion sizes.

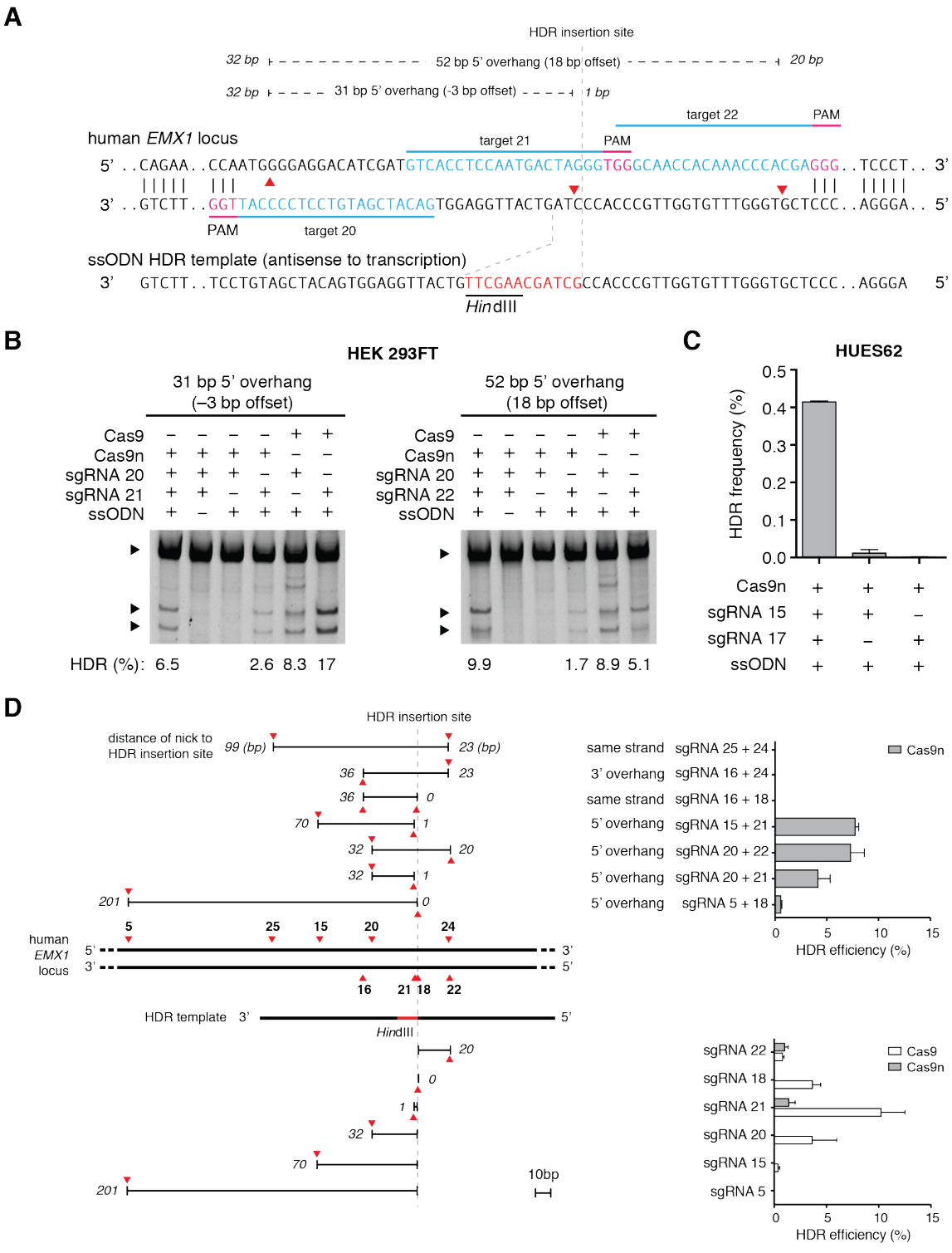


Figure 3-4 Double nicking allows insertion into the genome via HDR in human cells

(A) Schematic illustrating HDR mediated via a single stranded oligodeoxynucleotide (ssODN) template at a DSB created by a pair of Cas9n enzymes. A 12-nt sequence (red), including a *Hin*dIII restriction site, is inserted into the *EMX1* locus at the position marked by the gray dashed lines; distances of Cas9n-mediated nicks from the HDR insertion site is indicated on top in italics. (B) Restriction digest assay gel showing successful insertion of *Hin*dIII cleavage sites by double nicking-mediated HDR in HEK 293FT cells. Upper bands are unmodified template; lower bands are *Hin*dIII cleavage product. (C) Double nicking promotes HDR in the HUES62 human embryonic stem cell line. HDR frequencies are determined by deep sequencing. (n = 3; error bars show mean ± s.e.m.). (D) HDR efficiency depends on the configuration of Cas9 or Cas9n-mediated nicks. HDR is facilitated when a nick occurs near the center of the ssODN homology arm (HDR insertion site) leading to a 5'-resulting overhang. Nicking configurations are annotated with position and strand (red arrows) and length of overhang (black lines) (left panel). The distance (bp) of each nick from the HDR insertion site is indicated at the end of the black lines in italics, and the positions of the sgRNAs are illustrated in bold on the schematic of the *EMX1* locus. HDR efficiency mediated by double nicking with paired sgRNAs (top panel) or single sgRNAs with either Cas9 or Cas9n are shown (bottom panel, Supplementary Table 3; n = 3, error bars show mean ± s.e.m.).

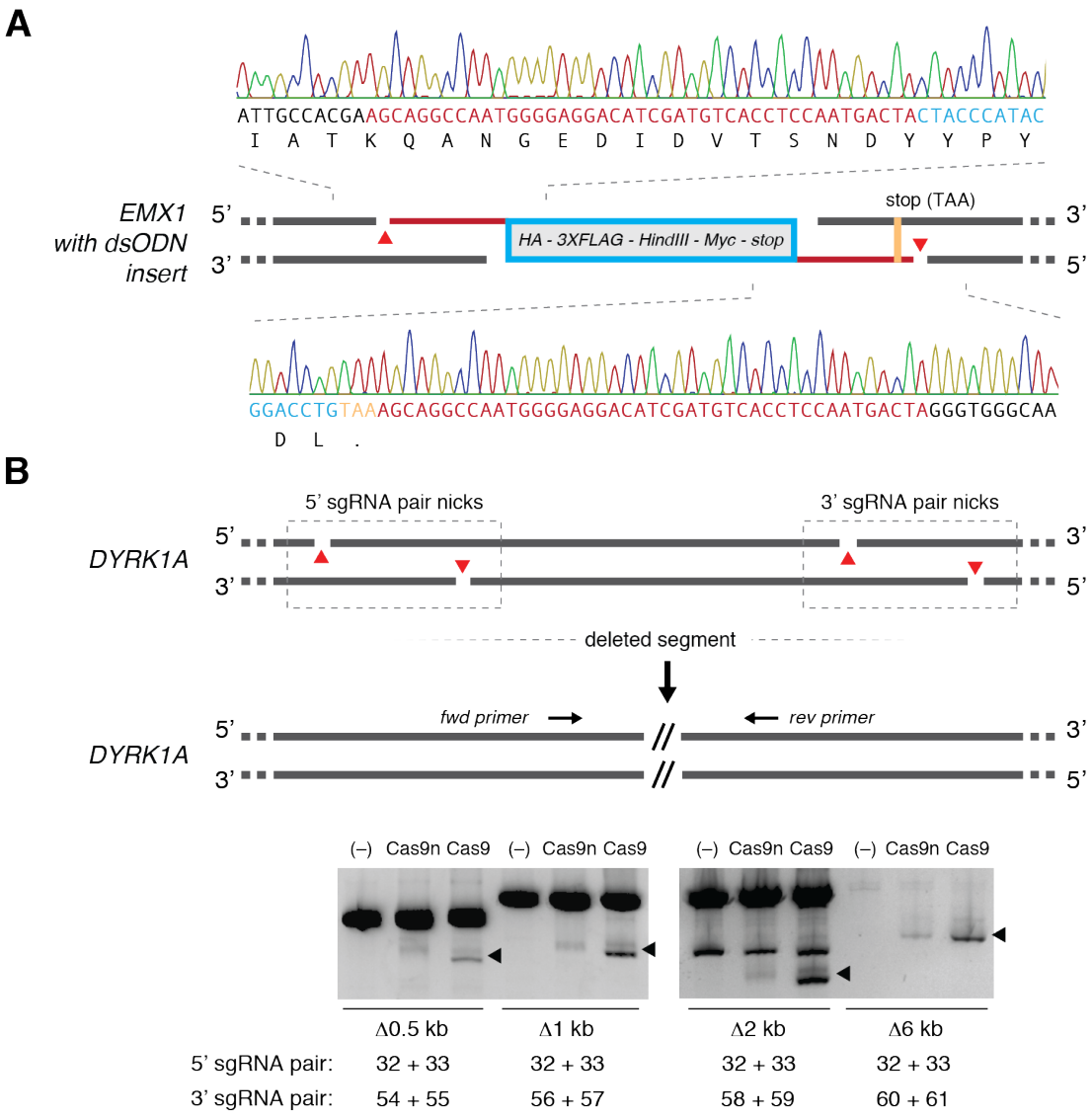


Figure 3-5 Multiplexed nicking facilitates non-HR mediated gene integration and genomic deletion

(A) Schematic showing insertion of a double-stranded oligodeoxynucleotide (dsODN) donor fragment bearing overhangs complementary to 5' overhangs created by Cas9 double nicking. The dsODN was designed to remove the native *EMX1* stop codon and contains a HA tag, 3X FLAG tag, *Hind*III restriction site, *Myc* epitope tag, and a stop codon in frame, totaling 148 bp. Successful insertion was verified by Sanger sequencing as shown (1/37 clones screened). Amino acid translation of the modified locus is shown below the DNA sequence. (B) Co-delivery of four sgRNAs with Cas9n generate long-range genomic deletions in the DYRK1A locus (from 0.5 kb up to 6 kb). Deletion was detected using primers spanning the target region.

Efficient genome modification in mouse zygotes

Recent work demonstrated that co-delivery of wildtype Cas9 mRNA along with multiple sgRNAs can mediate single-step generation of transgenic mice carrying multiple allelic modifications (*6*). Given the ability to achieve genome modification *in vivo* using several sgRNAs at once, we sought to assess the efficiency of multiple nicking by Cas9n in mouse zygotes. Cytoplasmic co-injection of wildtype Cas9 or Cas9n mRNA and sgRNAs into single-cell mouse zygotes allowed successful targeting of the *Mecp2* locus (Figure 3-6A). To identify the optimal concentration of Cas9n mRNA and sgRNA for efficient gene targeting, we titrated Cas9 mRNA from 100 ng/uL to 3 ng/uL while maintaining the sgRNA levels at a 1:20 Cas9:sgRNA molar ratio. All concentrations tested for Cas9 double nicking mediated modifications in at least 80% of embryos screened, similar to levels achieved by wildtype Cas9 (Figure 3-6B). Taken together, these results suggest a number of applications for double nicking-based genome editing.

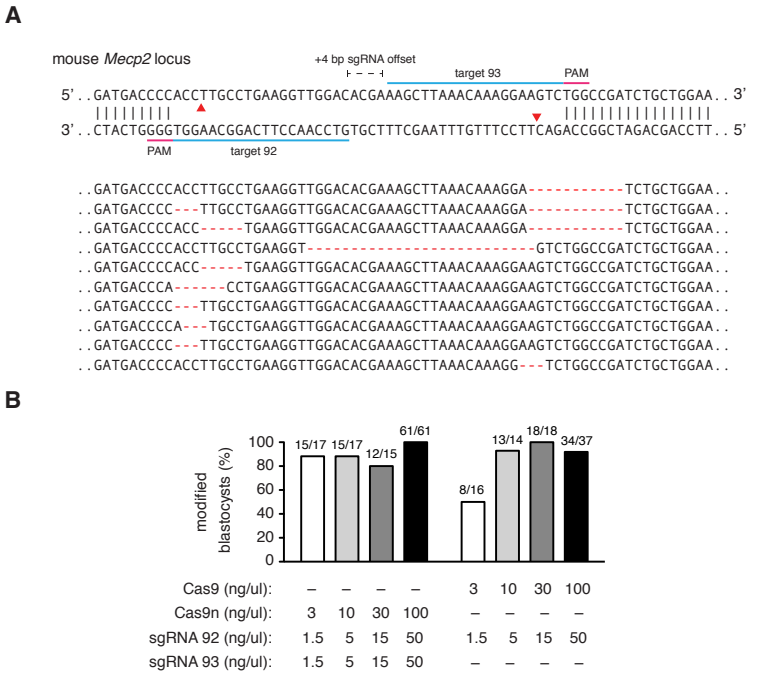


Figure 3-6 Cas9 double nicking mediates efficient indel formation in mouse embryos

(A) Schematic illustrating Cas9n double nicking the mouse *Mecp2* locus. Representative indels are shown for mouse blastocysts co-injected with *in vitro* transcribed Cas9n-encoding mRNA and sgRNA pairs matching targets 92 and 93. (B) Efficient blastocyst modification is achieved at multiple concentrations of sgRNAs (1.5 to 50 ng/uL) and wildtype Cas9 or Cas9n (ng/uL to 100 ng/uL).

Discussion

Given the permanent nature of genomic modifications, specificity is of paramount importance to sensitive applications such as studies aimed at linking specific genetic variants with biological processes or disease phenotypes and gene therapy. Here, we have explored strategies to improve the targeting specificity of Cas9. Although simply extending the guide sequence length of sgRNA failed to improve targeting specificity, combining two appropriately offset sgRNAs with Cas9n effectively generated indels while minimizing unwanted cleavage since individual off-target single-stranded nicks are repaired with high fidelity via base excision repair. Given that significant off-target mutagenesis has been previously reported for Cas9 nucleases in human cells (*21, 22*), the DN approach could provide a generalizable solution for rapid and accurate genome editing.

The characterization of spacing parameters governing successful Cas9 double nickase-mediated gene targeting reveals an effective offset window over 100-bp long, allowing for a high degree of flexibility in the selection of sgRNA pairs. Previous computational analyses have revealed an average targeting range of every 12-bp for the *Streptococcus pyogenes* Cas9 in the human genome based on the 5'-NGG PAM (*5*), which suggest that appropriate sgRNA pairs should be readily identifiable for most loci within the genome. We have additionally demonstrated DN-mediated indel frequencies comparable to wild type Cas9 modification at multiple genes and loci in both human and mouse cells, confirming the reproducibility of this strategy for high-precision genome engineering (Supplementary Table 2).

The Cas9 double nicking approach is in principle similar to ZFN and TALEN-based genome editing systems, where cooperation between two hemi-nuclease domains is required to achieve double-stranded break at the target site. Systematic studies of ZFN and TALEN systems have revealed that the targeting specificity of a given ZFN

and TALEN pair can be highly dependent on the nuclease architecture (homo- or heterodimeric nucleases) or target sequence, and in some cases TALENs can be highly specific (*40*). Although the wildtype Cas9 system has been shown to exhibit high levels of off-target mutagenesis, the DN system is a promising solution and brings RNA-guided genome editing to similar specificity levels as ZFNs and TALENs.

Additionally, the ease and efficiency with which Cas9 can be targeted renders the DN system especially attractive. However, DNA targeting using DN will likely face similar off-target challenges as ZFNs and TALENs, where cooperative nicking at off-target sites might still occur, albeit at a significantly reduced likelihood. Given the extensive characterization of Cas9 specificity and sgRNA mutation analysis (*21, 22*), as well as the NHEJ-mediating sgRNA offset range identified in this study, computational approaches may be used to evaluate the likely off-target sites for a given pair of sgRNAs. To facilitate sgRNA pair selection, we developed an online web tool that identifies sgRNA combinations with optimal spacing for double nicking applications (http://www.genome-engineering.org/).

Although Cas9n has been previously shown to facilitate HDR at on-target sites (*5*), its efficiency is substantially lower than that of wildtype Cas9. The double nicking strategy, by comparison, maintains high on-target efficiencies while reducing off-target modifications to background levels. Nevertheless, further characterizations of DN off-target activity, particularly via whole genome sequencing and targeted deep sequencing of cells or whole organisms generated using the DN approach, are urgently needed to evaluate the utility of Cas9n DN in biotechnological or clinical applications that require ultra-high precision genome editing. Additionally, Cas9n has been shown to induce low levels of indels at on-target sites for certain sgRNAs (*7*), which may result from residual double-strand break activities and be circumvented by further structure-function studies of Cas9 catalytic activity. Overall, Cas9n-mediated multiplex nicking serves as a customizable platform for highly precise and efficient targeted genome engineering and promises to broaden the range of applications in biotechnology, basic science, and medicine.

References

1. D. F. Carlson *et al.*, Efficient TALEN-mediated gene knockout in livestock. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17382 (Oct 23, 2012).

2. A. Geurts *et al.*, Knockout rats via embryo microinjection of zinc-finger nucleases. *Science (New York, N.Y.)* **325**, 433 (2009).

3. Y. Takasu *et al.*, Targeted mutagenesis in the silkworm Bombyx mori using zinc finger nuclease mRNA injection. *Insect biochemistry and molecular biology* **40**, 759 (Oct, 2010).

4. T. Watanabe *et al.*, Non-transgenic genome modifications in a hemimetabolous insect using zinc-finger and TAL effector nucleases. *Nat Commun* **3**, 1017 (2012).

5. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

6. H. Wang *et al.*, One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910 (May 9, 2013).

7. P. Mali *et al.*, RNA-guided human genome engineering via Cas9. *Science* **339**, 823 (Feb 15, 2013).

8. M. Jinek *et al.*, RNA-programmed genome editing in human cells. *eLife* **2**, (2013).

9. H. Deveau, J. Garneau, S. Moineau, CRISPR/Cas system and its role in phage-bacteria interactions. *Annual review of microbiology* **64**, 475 (2010).

10. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602 (2011).

11. N. Chang *et al.*, Genome editing with RNA-guided Cas9 nuclease in zebrafish embryos. *Cell research* **23**, 465 (Apr, 2013).

12. S. Gratz *et al.*, Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics* **194**, 1029 (2013).

13. A. Friedland *et al.*, Heritable genome editing in C. elegans via a CRISPR-Cas9 system. *Nature methods* **10**, 741 (2013).

14. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167 (Jan 8, 2010).

15. S. W. Cho, S. Kim, J. M. Kim, J. S. Kim, Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature biotechnology* **31**, 230 (Mar, 2013).

16. E. Perez *et al.*, Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature biotechnology* **26**, 808 (2008).

17. F. Urnov, E. Rebar, M. Holmes, H. Zhang, P. Gregory, Genome editing with engineered zinc finger nucleases. *Nature reviews. Genetics* **11**, 636 (2010).

18. P. D. Hsu, F. Zhang, Dissecting neural function using targeted genome engineering technologies. *ACS chemical neuroscience* **3**, 603 (Aug 15, 2012).

19. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816 (2012).

20. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579 (Sep 25, 2012).

21. Y. Fu *et al.*, High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature biotechnology* **31**, 822 (Sep, 2013).

22. P. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827 (2013).

23. W. Jiang, D. Bikard, D. Cox, F. Zhang, L. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* **31**, 233 (2013).

24. F. Soldner *et al.*, Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* **146**, 318 (2011).

25. G. L. Dianov, U. Hubscher, Mammalian base excision repair: the forgotten archangel. *Nucleic acids research* **41**, 3483 (Apr 1, 2013).

26. P. Mali *et al.*, CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**, 833 (2013).

27. M. H. Porteus, D. Baltimore, Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (May 2, 2003).

28. J. Miller *et al.*, An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature biotechnology* **25**, 778 (2007).

29. J. Sander *et al.*, Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nature methods* **8**, 67 (2011).

30. A. J. Wood *et al.*, Targeted genome editing across species using ZFNs and TALENs. *Science* **333**, 307 (Jul 15, 2011).

31. M. Christian *et al.*, Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757 (Oct, 2010).

32. D. Hockemeyer *et al.*, Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology* **29**, 731 (Aug, 2011).

33. D. Reyon *et al.*, FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology* **30**, 460 (2012).

34. J. Boch *et al.*, Breaking the code of DNA binding specificity of TAL-type III effectors. *Science (New York, N.Y.)* **326**, 1509 (2009).

35. M. Moscou, A. Bogdanove, A simple cipher governs DNA recognition by TAL effectors. *Science (New York, N.Y.)* **326**, 1501 (2009).

36. N. E. Sanjana *et al.*, A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* **7**, 171 (Jan, 2012).

37. P. Mali *et al.*, CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**, 833 (Sep, 2013).

38. V. Pattanayak *et al.*, High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature biotechnology* **31**, 839 (Sep, 2013).

39. M. Maresca, V. Lin, N. Guo, Y. Yang, Obligate ligation-gated recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome research* **23**, 539 (2013).

40. Q. Ding *et al.*, A TALEN genome-editing system for generating human stem cell-based disease models. *Cell stem cell* **12**, 238 (Feb 7, 2013).

Chapter 4 Crystal structure of Cas9 in complex with guide RNA and target DNA

Hiroshi Nishimasu[1,2], F. Ann Ran[3,4,5,6,7], Patrick D. Hsu[3,4,5,6,7], Silvana Konermann[3,4,5,6], Soraya Shehata[3,4,5,6], Naoshi Dohmae[8], Ryuichiro Ishitani[1], Feng Zhang[3,4,5,6], and Osamu Nureki[1]

[1] Department of Biophysics and Biochemistry
Graduate School of Science
The University of Tokyo
[2] JST, PRESTO
2-11-16 Yayoi, Bunkyo, Tokyo, 113-0032, Japan

[3] Broad Institute of MIT and Harvard
[4] McGovern Institute for Brain Research
[5] Department of Brain and Cognitive Sciences
[6] Department of Biological Engineering
Massachusetts Institute of Technology
Cambridge, MA, USA

[7] Department of Molecular and Cellular Biology
Harvard University
Cambridge, MA, USA

[8] Biomolecular Characterization Team and CREST/JST
RIKEN, 2-1 Hirosawa
Wako, Saitama 351-0198, Japan

Acknowledgements

We thank Arisa Kurabayashi for assistance with vector construction, Tomohiro Nishizawa and Kazuki Kato for helpful comments on the manuscript, Ryohei Ishii and Motoyuki Hattori for assistance in data collection, and Chie-Yu Lin for help with the sgRNA mutagenesis analysis. We thank the beam-line staffs at BL32XU and BL41XU of SPring-8, Japan, and at IO3 of Diamond Light Source, UK, for assistance with data collection. F.Z. is supported by an NIH Director's Pioneer Award (5DP1- MH100706); by the Keck, McKnight, Poitras, Merkin, Vallee, Damon Runyon, Searle Scholars, Klingenstein, and Simons Foundations; and by Bob Metcalfe and Jane Pauley.

Author Contributions

H.N. performed the crystallization and structural analysis; F.A.R. designed and performed the Cas9 truncation, sgRNA mutation, and nickase analyses; P.D.H., S.K., and S.I.S. designed and performed the Cas9 domain switching and point mutation analyses; N.D. performed the mass spectrometric analysis; R.I. performed the structural analysis; and H.N., F.A.R., P.D.H., F.Z., and O.N. wrote the manuscript with help from all authors. H.N., F.Z., and O.N. directed and supervised all of the research.

This chapter contains work from the manuscript "Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA," published in *Cell*, February 27, 2014, Vol. 156 pp. 935–949. The text and figures were modified to fit the format of this dissertation.

Introduction

The CRISPR (clustered regularly interspaced palindromic repeat)-Cas system is a naturally occurring, adaptive microbial immune system for defense against invading phages and other mobile genetic elements (*1-4*). Three types (I–III) of CRISPR-Cas systems have been functionally identified across a wide range of microbial species (*5-7*), and each contains a cluster of CRISPR-associated (*Cas*) genes and its corresponding CRISPR array. These characteristic CRISPR arrays consist of repetitive sequences (direct repeats, referred to as repeats) interspaced by short stretches of non-repetitive sequences (spacers) derived from short segments of foreign genetic material (protospacers). The CRISPR array is transcribed and processed into short CRISPR RNAs (crRNAs), which direct the Cas proteins to the target nucleic acids via Watson-Crick base pairing to facilitate nucleic acid destruction.

The Type I and III CRISPR systems utilize ensembles of Cas proteins complexed with crRNAs to mediate the recognition and subsequent degradation of target nucleic acids (*8, 9*). In contrast, the Type II CRISPR system recognizes and cleaves the target DNA (*10*) via the RNA-guided endonuclease Cas9 (*11*) along with two non-coding RNAs, the crRNA and the trans-activating crRNA (tracrRNA) (*13*). The crRNA hybridizes with the tracrRNA to form a crRNA:tracrRNA duplex, which is loaded onto Cas9 to direct the cleavage of cognate DNA sequences bearing appropriate protospacer adjacent motifs (PAM) (*15*). Cas9 contains two nuclease domains, HNH and RuvC, which cleave the DNA strands that are complementary and non-complementary to the 20-nucleotide (nt) guide sequence in crRNAs, respectively (*16, 17*).

The Type II CRISPR system was the first to be adapted for facilitating genome editing in eukaryotic cells (*18, 19*). The Cas9 protein from *Streptococcus pyogenes*, along with a single guide RNA (sgRNA), a synthetic fusion of crRNA and tracrRNA (*16*), could be programmed to cleave virtually any sequence preceding a 5′-NGG-3′ PAM sequence in mammalian cells (*18, 19*). This unprecedented flexibility has enabled a broad range of applications,

including rapid generation of genetically modified cells and animal models (*20-23*), and genome-scale genetic screening (*24, 25*).

However, despite the brisk progress in the development of Cas9 technology, the mechanism by which the Cas9–sgRNA complex recognizes and cleaves its target DNA remains to be elucidated. Biochemical analyses at the domain levels have enabled site-specific engineering to convert the native Cas9 into a DNA nicking enzyme (*11, 16, 17*) that facilitates homology-directed repair in eukaryotic cells (Cong et al., 2013; Mali et al., 2013b) and cleaves DNA with improved specificity, given appropriately paired sgRNAs (*26, 27*). Moreover, a catalytically inactive Cas9 can serve as an RNA-guided DNA-binding platform to target effector domains and modulate endogenous transcription (*28-32*). These advances in Cas9 engineering represent just the first steps toward fully realizing the potential of this flexible RNA-guided genome positioning system. Precise structural information about Cas9 will thus not only enhance our understanding of how this elegant RNA-guided, adaptive microbial immune system functions, but also facilitate further improvements in the Cas9 targeting specificity, the *in vitro* and *in vivo* delivery, and the engineering of Cas9 for novel functions and optimized features.

Here, we report the crystal structure of *S. pyogenes* Cas9 in complex with sgRNA and its target DNA at 2.5 Å resolution. This high-resolution structure, along with functional analyses, reveals the key functional interactions that integrate the guide RNA, the target DNA, and the Cas9 protein, thus paving the way towards enhancing Cas9 function as well as engineering novel applications.

Overall structure of the Cas9–sgRNA–DNA ternary complex

We solved the crystal structure of full-length *S. pyogenes* Cas9 (residues 1–1368; D10A/C80L/C574E/H840A), in complex with a 98-nt sgRNA and a 23-nt target DNA, at 2.5 Å resolution by the SAD (single-wavelength anomalous dispersion) method, using a SeMet-labeled protein (Figure 4-1, Supplementary Figure 10, and Supplementary Table 4). To improve the solution behavior of Cas9, we replaced two less conserved cysteine residues (Cys80 and Cys574) with leucine and glutamic acid, respectively. This C80L/C574E mutant retained the ability to efficiently cleave genomic DNA in human embryonic kidney 293FT (HEK293FT) cells, confirming that these mutations have no effects on the Cas9 nuclease function (Supplementary Figure 11). Additionally, to prevent target DNA cleavage during crystallization, we replaced two catalytic residues, Asp10 from the RuvC domain and His840 from the HNH domain, with alanines.

The crystallographic asymmetric unit contained two Cas9–sgRNA–DNA ternary complexes (Mol A and Mol B). Although there are conformational differences between the two complexes, the sgRNA and the DNA are recognized by Cas9 in similar manners. Most notably, while the HNH domain in Mol A is connected to the RuvC domain by a disordered linker, the HNH domain in Mol B is not visible in the electron density map, indicating the flexible nature of the HNH domain. Thus, we will first describe the structural features of Mol A unless otherwise stated, and then discuss the structural differences between the two complexes, which suggest the conformational flexibility of Cas9.

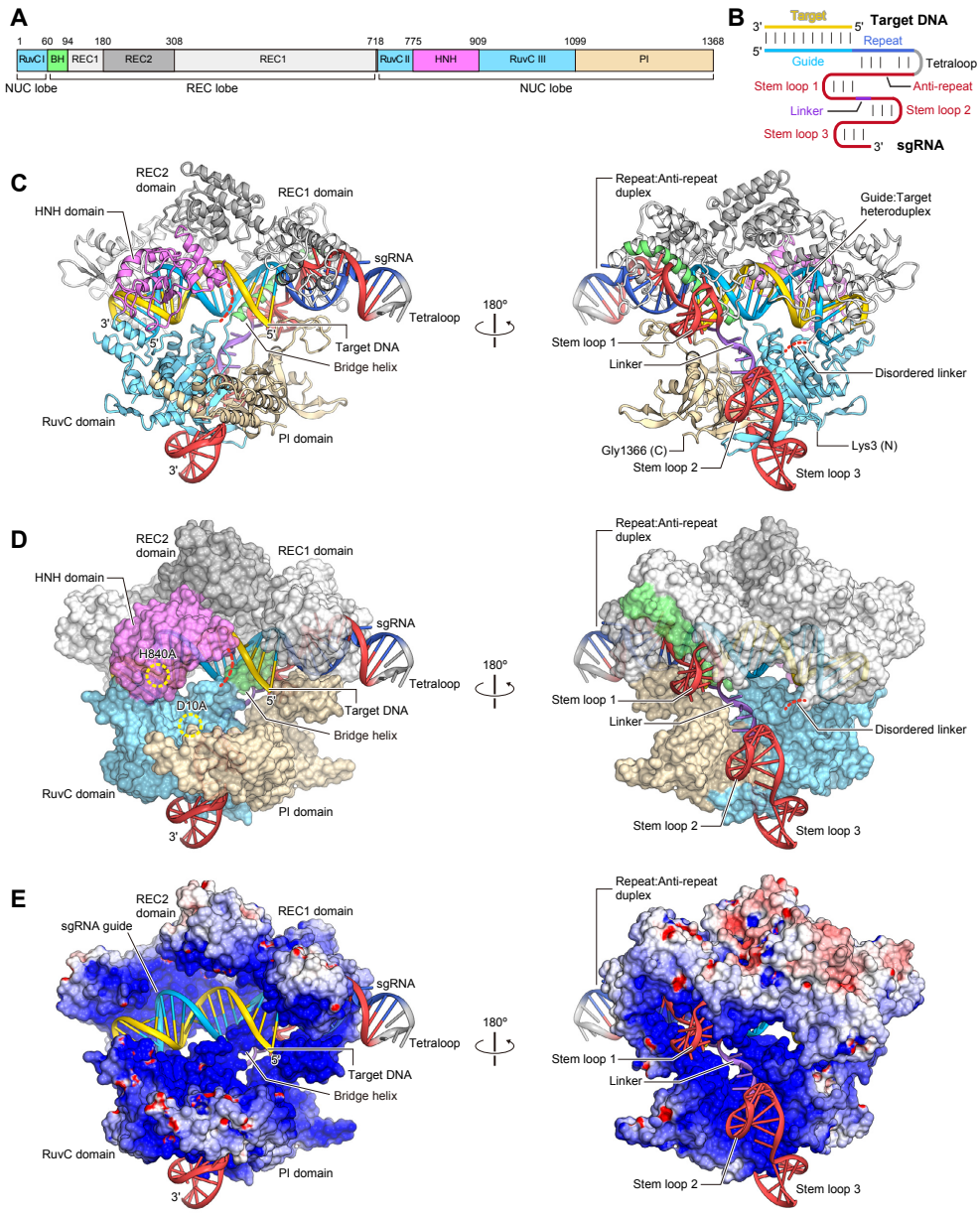


Figure 4-1 Overall structure of the Cas9–sgRNA–DNA ternary complex.

(A) Domain organization of *S. pyogenes* Cas9. BH, Bridge helix. (B) Schematic representation of the sgRNA:target DNA complex. (C) Ribbon representation of the Cas9–sgRNA–DNA complex. Disordered linkers are shown as red dotted lines. (D) Surface representation of the Cas9–sgRNA–DNA complex. The active sites of the RuvC (D10A) and HNH (H840A) domains are indicated by dashed yellow circles. (E) Electrostatic surface potential of Cas9. The HNH domain is omitted for clarity. Molecular graphic images were prepared using CueMol (http://www.cuemol.org).

The crystal structure revealed that Cas9 consists of two lobes, a recognition (REC) lobe and a nuclease (NUC) lobe (Figure 4-1A–D). The REC lobe can be divided into three regions, a long α-helix referred to as the Bridge helix (residues 60–93), the REC1 (residues 94–179 and 308–713) domain, and the REC2 (residues 180–307) domain (Figure 4-1A–D). The NUC lobe consists of the RuvC (residues 1–59, 718–769 and 909–1098), HNH (residues 775–908), and PAM-interacting (PI) (residues 1099–1368) domains (Figure 4-1A–D). The negatively-charged sgRNA:target DNA heteroduplex is accommodated in a positively-charged groove at the interface between the REC and NUC lobes (Figure 4-1E). In the NUC lobe, the RuvC domain is assembled from the three split RuvC motifs (RuvC I–III), and interfaces with the PI domain to form a positively-charged surface that interacts with the 3′ tail of the sgRNA (Figure 4-1E). The HNH domain lies in between the RuvC II–III motifs and forms only a few contacts with the rest of the protein.

The REC lobe interacts with the repeat:anti-repeat duplex

The REC lobe includes the REC1 and REC2 domains. REC1 adopts an elongated, α-helical structure comprising 25 α-helices (α2–α5 and α12–α32) and two β-sheets (β6 and β10, and β7–β9), whereas REC2 adopts a six-helix bundle structure (α6–α11) (Figure 4-2A and Supplementary Figure 12). A Dali search (*33*) revealed that the REC lobe does not share structural similarity with other known proteins, indicating that it is a Cas9-specific functional domain.

The REC lobe is one of the least conserved regions across the three Cas9 families within the Type II CRISPR system (IIA, IIB and IIC), and many Cas9 orthologs contain significantly shorter REC lobes (Supplementary Figure 13 and Supplementary Figure 14). In the present structure, the REC2 domain does not contact the bound guide:target heteroduplex. We thus hypothesized that truncations in the REC lobe could be tolerated. As expected, a Cas9 mutant lacking the REC2 domain (Δ175–307) retained ~50% of the wild-type Cas9 activity (Figure 4-2B), indicating that the REC2 domain is not critical for DNA cleavage. The lower cleavage efficiency

may be attributed in part to the reduced expression levels of the Δ175–307 mutant relative to that of the wild-type protein (Figure 4-2C). In striking contrast, the deletion of either the repeat-interacting region (Δ97–150) or the anti-repeat-interacting region (Δ312–409) of the REC1 domain abolished the DNA cleavage activity (Figure 4-2B), indicating that the recognition of the repeat:anti-repeat duplex by the REC1 domain is critical for the Cas9 function.

The PAM-interacting (PI) domain confers PAM specificity

The NUC lobe contains the PI domain, which forms an elongated structure comprising seven α-helices (α46–α52), a three-stranded antiparallel β-sheet (β18–β20), a five-stranded antiparallel β-sheet (β21–β23, β26 and β27), and a two-stranded antiparallel β-sheet (β24 and β25) (Figure 4-2D and Supplementary Figure 12). Similar to the REC lobe, the PI domain also adopts a novel protein fold unique to the Cas9 family.

The locations of the bound complementary DNA strand and the active site of the RuvC domain in the present structure suggested that the PI domain is positioned to recognize the PAM sequence on the non-complementary DNA strand. We tested whether the replacement of the PI domain of *S. pyogenes* Cas9 (SpCas9; Cas9 in this study) with that of an orthologous Cas9 protein, recognizing a different PAM, would be sufficient to alter their PAM specificities. The *Streptococcus thermophilus* CRISPR-3 Cas9 (St3Cas9) shares ~60% sequence identity with SpCas9. While their crRNA repeats and tracrRNAs are interchangeable (*34*), SpCas9 and St3Cas9 require different PAM sequences (5′-N<u>GG</u>-3′ for SpCas9 and 5′-N<u>GG</u>N<u>G</u>-3′ for St3Cas9) for target DNA cleavage (*34*).

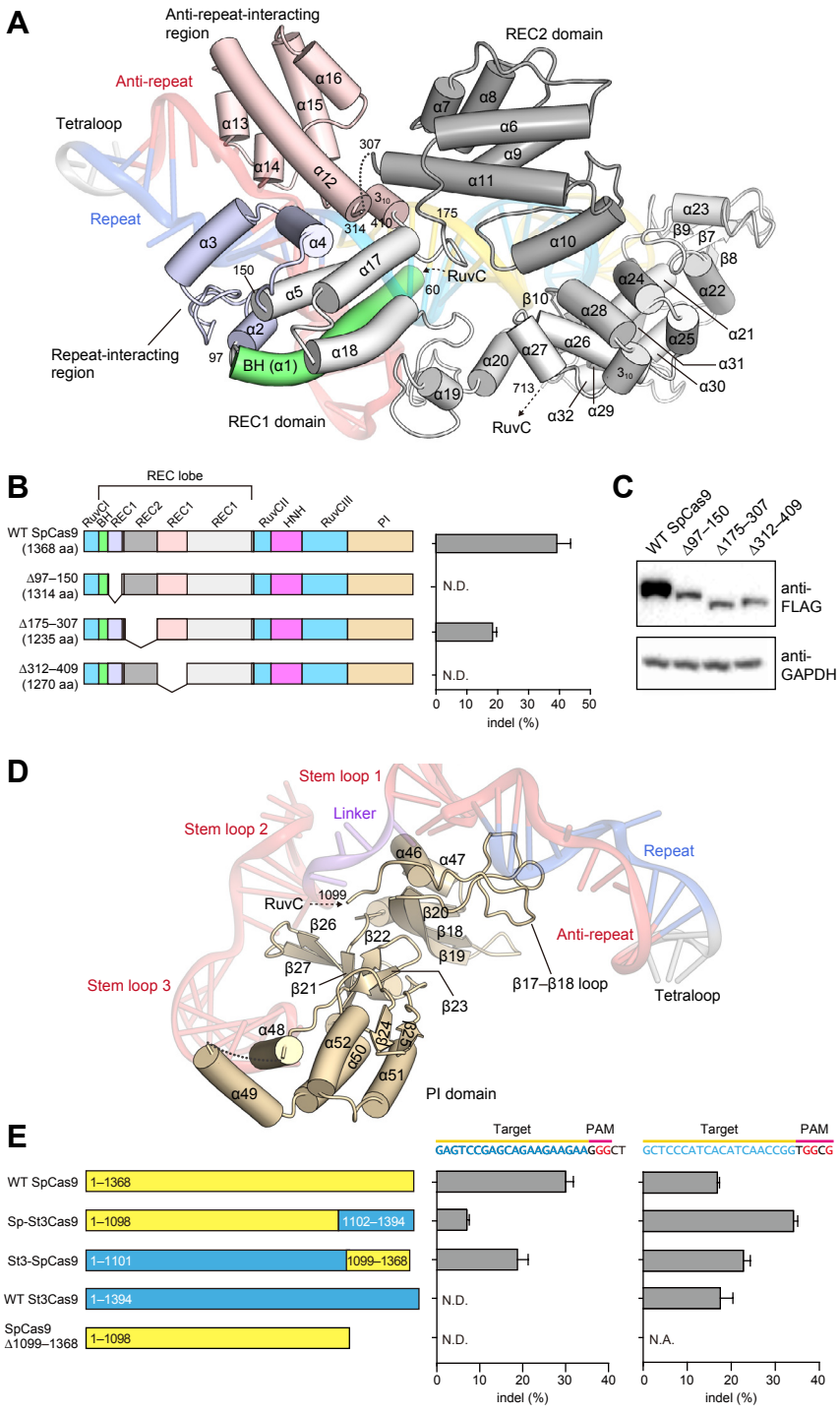


Figure 4-2 REC lobe and PI domain

(A) Structure of the REC lobe. The REC2 domain and the Bridge helix are colored dark gray and green, respectively. The REC1 domain is colored gray, with the repeat-interacting and anti-repeat-interacting regions colored pale blue and pink, respectively. The bound sgRNA:DNA is shown as a semi-transparent ribbon representation.
(B) Mutational analysis of the REC lobe. Schematics show the truncation mutants. The bar graph shows indel mutations generated by the truncation mutants, measured by the SURVEYOR assay (n = 3, error bars show mean ± S.E.M., N.D., not detectable).
(C) Western blot showing the expression of the truncation mutants in HEK 293FT cells.
(D) Structure of the PI domain. The bound sgRNA is shown as a semi-transparent ribbon representation.
(E) Mutational analysis of the PI domain. Schematics show wild-type SpCas9 and St3Cas9, chimeric Sp-St3Cas9 and St3-SpCas9, and the SpCas9 PI domain truncation mutant. Cas9s were assayed for indel generation at target sites upstream of either NGG (left bar graph) or NGGNG (right bar graph) PAMs (n = 3, error bars show mean ± S.E.M., N.D., not detectable).

We swapped their PI domains to generate two chimeras, Sp-St3Cas9 (SpCas9 with the PI domain of St3Cas9) and St3-SpCas9 (St3Cas9 with the PI domain of SpCas9), and examined their cleavage activities for target DNA sequences bearing 5′-NGG-3′ PAM (5′-GGGCT-3′) or 5′-NGGNG-3′ PAM (5′-GGGCG-3′) (Figure 4-2E). SpCas9 and St3-SpCas9, but not St3Cas9, cleaved the target DNA with 5′-NGG-3′ PAM (Figure 4-2E), indicating that the PI domain of SpCas9 is required for the recognition of 5′-NGG-3′ PAM and is sufficient to alter the PAM specificity of St3Cas9. Sp-St3Cas9 retained the cleavage activity for the target DNA with 5′-NGG-3′ PAM, albeit at a lower level than that of SpCas9 (Figure 4-2E). Additionally, the deletion of the PI domain (Δ1099–1368) abolished the cleavage activity (Figure 4-2E), indicating that the PI domain is critical for the Cas9 function. These results revealed that the PI domain is a major determinant of the PAM specificity.

The RuvC domain has an RNase H fold

The RuvC domain consists of a six-stranded mixed β-sheet (β1, β2, β5, β11, β14 and β17) flanked by α-helices (α33, α34 and α39–α45) and two additional two-stranded antiparallel β-sheets (β3/β4 and β15/β16) (Figure 4-3A and Supplementary Figure 12). It shares structural similarity with the retroviral integrase superfamily members characterized by an RNase H fold, such as *Escherichia coli* RuvC (35) (PDB code 1HJR, 14% identity, root-mean-square deviation (rmsd) of 3.6 Å for 126 equivalent Cα atoms) and *Thermus thermophilus* RuvC (Gorecka et al., 2013) (PDB code 4LD0, 12% identity, rmsd of 3.4 Å for 131 equivalent Cα atoms) (Figure 4-3B). The RuvC nucleases have four catalytic residues (e.g., Asp7, Glu70, His143 and Asp146 in *T. thermophilus* RuvC), and cleave Holliday junctions through a two-metal mechanism (35-37). Asp10 (Ala), Glu762, His983 and Asp986 of the Cas9 RuvC domain are located at positions similar to those of the catalytic residues of *T. thermophilus* RuvC (Figure 4-3A, B), consistent with previous results showing that Asp10 is critical for the cleavage of the non-complementary DNA strand, and that Cas9 requires Mg^{2+} ions for the cleavage activity (*16, 17*). Moreover, the alanine substitution of Glu762, His983 or Asp986 also converted Cas9 into a nickase (Figure 4-3C, D). Each nickase mutant was able to facilitate targeted double strand breaks using a pair of juxtaposed

sgRNAs (Figure 4-3C, D, Supplementary Table 5), as previously demonstrated with the D10A nickase (*26*). This combination of structural observations and mutational analyses suggested that the Cas9 RuvC domain cleaves the non-complementary strand of the target DNA through the two-metal mechanism, as previously observed for other retroviral integrase superfamily nucleases.

It is important to note that there are key structural dissimilarities between the Cas9 RuvC domain and the RuvC nucleases, which explain their functional differences. Unlike the Cas9 RuvC domain, the RuvC nucleases form dimers and recognize Holliday junctions (*36*) (Figure 4-3B). In addition to the conserved RNase H fold, the Cas9 RuvC domain has other structural elements involved in interactions with the guide:target heteroduplex (an end-capping loop between α42 and α43) and the PI domain/stem loop 3 (β-hairpin formed by β3 and β4) (Figure 4-3A).

The HNH domain has a ββα-metal fold

The HNH domain comprises a two-stranded antiparallel β-sheet (β12 and β13) flanked by four α-helices (α35–α38) (Figure 4-3E). It shares structural similarity with the HNH endonucleases characterized by a ββα-metal fold, such as phage T4 endonuclease VII (Endo VII) (*38*) (PDB code 2QNC, 20% identity, rmsd of 2.7 Å for 61 equivalent Cα atoms) and *Vibrio vulnificus* nuclease (*39*) (PDB code 1OUP, 8% identity, rmsd of 2.7 Å for 77 equivalent Cα atoms) (Figure 4-3F). HNH nucleases have three catalytic residues (e.g., Asp40, His41, and Asn62 in Endo VII), and cleave nucleic acid substrates through a single-metal mechanism (*38, 39*). In the structure of the Endo VII N62D mutant in complex with a Holliday junction, a Mg^{2+} ion is coordinated by Asp40, Asp62, and the oxygen atoms of the scissile phosphate group of the substrate, while His41 acts as a general base to activate a water molecule for catalysis (Figure 4-3F). Asp839, His840, and Asn863 of the Cas9 HNH domain correspond to Asp40, His41, and Asn62 of Endo VII, respectively (Figure 4-3E), consistent with the observation that His840 is critical for the cleavage of the complementary DNA strand (*16, 17*). The N863A mutant functions as a nickase

(Figure 4-3C, D), indicating that Asn863 participates in catalysis. These observations suggested that the Cas9 HNH domain may cleave the complementary strand of the target DNA through a single-metal mechanism, as observed for other HNH superfamily nucleases. However, in the present structure, Asn863 of Cas9 is located at a different position from that of Asn62 in Endo VII, whereas Asp839 and His840 (Ala) of Cas9 are located at positions similar to those of Asp40 and His41 in Endo VII, respectively (Figure 4-3G). This might be due to the absence of divalent ions, such as Mg^{2+}, in our crystallization solution, and Asn863 may point towards the active site and participate in catalysis. Although the HNH domain shares a ββα-metal fold with other HNN endonucleases, their overall structures are distinct (Figure 4-3E, F), consistent with the differences in their substrate specificities.

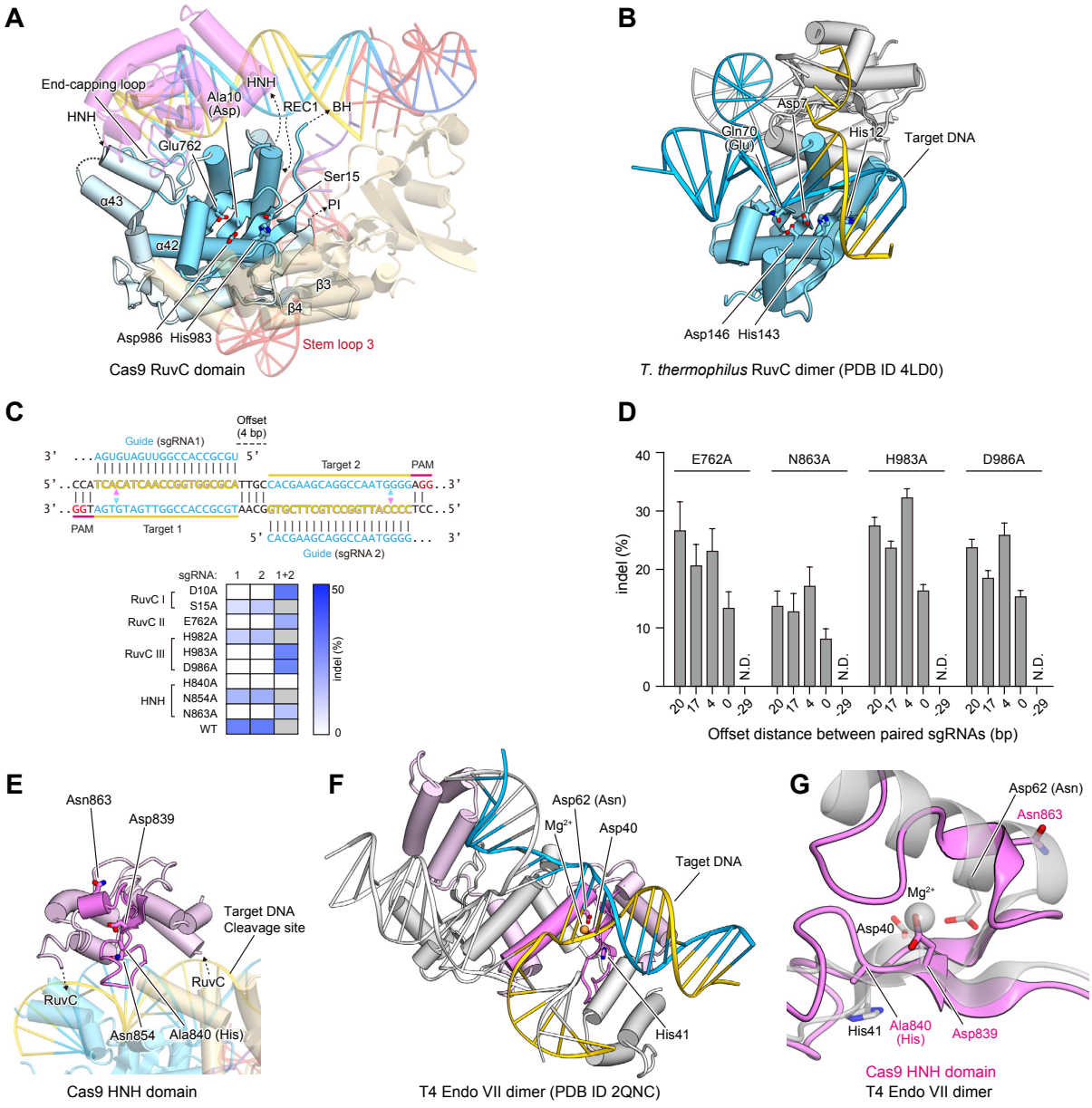


Figure 4-3 NUC lobe

(A) Structure of the RuvC domain. The core structure of the RNase H fold is highlighted in cyan. The active-site residues are shown as stick models. (B) Structure of the *T. thermophilus* RuvC dimer in complex with a Holliday junction (PDB ID 4LD0). The two protomers are colored cyan and gray, respectively. (C) Mutational analysis of the RuvC and HNH domains. The sequences (top) illustrate Cas9 nicking targets on opposite strands of DNA. Targets 1 and 2 are offset by a distance of 4-bp in between. The cleavage sites by the HNH and RuvC domains are indicated by pink and cyan triangles, respectively. The heatmap (bottom) shows the ability of each catalytic mutant to induce double- (with either sgRNA 1 or 2) or single-stranded breaks (only with both sgRNAs together). Gray boxes, not assayed. (D) Indel formation by Cas9 nickases depends on the off-set distance between sgRNA pairs. The off-set distance is defined as the number of base pairs between the PAM-distal (5′) ends of the guide sequence of a given sgRNA pair (n = 3, error bars show mean ± S.E.M., N.D., not detectable). (E) Structure of the HNH domain. The core structure of the bba-metal fold is highlighted in magenta. The active-site residues are shown as stick models. (F) Structure of the T4 Endo VII dimer in complex with a Holliday junction (PDB ID 2QNC). The two protomers are colored pink and gray, respectively, with the bba-metal fold core highlighted in magenta. The bound Mg^{2+} ion is shown as an orange sphere. (G) Superimposition of the Cas9 HNH domain and T4 Endo VII (PDB ID 2QNC).

The sgRNA:DNA complex adopts a T-shaped architecture

The sgRNA consists of crRNA- and tracrRNA-derived sequences connected by an artificial tetraloop (Figure 4-4A). The crRNA sequence can be divided into guide (20-nt) and repeat (12-nt) regions, while the tracrRNA sequence can be divided into anti-repeat (14-nt) and three tracrRNA stem loops (Figure 4-4A). The crystal structure revealed that the sgRNA binds the target DNA to form a T-shaped architecture comprising a guide:target heteroduplex, a repeat:anti-repeat duplex, and stem loops 1–3 (Figure 4-4A, B). The repeat:anti-repeat duplex and stem loop 1 are connected by a single nucleotide (A51), while stem loops 1 and 2 are connected by a 5-nt single-stranded linker (nucleotides 63–67).

The guide (nucleotides 1–20) and target DNA (nucleotides 1′–20′) form the guide:target heteroduplex via 20 Watson-Crick base pairs (Figure 4-4A, B). The conformation of the heteroduplex is distorted from that of the canonical A-form RNA duplex (Supplementary Figure 15). The repeat (nucleotides 21–32) and the anti-repeat (nucleotides 37–50) form the repeat:anti-repeat duplex via nine Watson-Crick base pairs (U22:A49–A26:U45 and G29:C40–A32:U37) (Figure 4-4A, B). Within this region, G27, A28, A41, A42, G43 and U44 are unpaired, with A28 and U44 flipped out from the duplex (Figure 4-4C). The G27 and A41 nucleobases stack with the A26:U45 and G29:C40 pairs, respectively, and the 2-amino group of G27 interacts with the backbone phosphate group between G43 and U44, stabilizing the duplex structure (Figure 4-4C). G21 and U50 form a wobble base pair at the three-way junction between the guide:target heteroduplex/repeat:anti-repeat duplexes and stem loop 1, contributing to the formation of the T-shaped architecture (Figure 4-4C).

As expected from the RNA-fold predictions based on the nucleotide sequence, the tracrRNA 3′ tail (nucleotides 68–81 and 82–96) forms stem loops 2 and 3 via four and six Watson-Crick base pairs (A69:U80–U72:A77 and G82:C96–G87:C91), respectively (Figure 4-4A, B). In addition, nucleotides 52–62 form the newly detected stem loop (stem loop 1) via three Watson-Crick base pairs (G53:C61, G54:C60 and C55:G58), with U59 flipped out

from the stem (Figure 4-4A, B). Stem loop 1 is stabilized by the G62–G53:C61 stacking interaction and the G62–A51/A52 polar interactions (Figure 4-4C).

The guide:target and repeat:anti-repeat duplexes are deeply buried in a positively-charged groove at the interface of the two lobes, while the rest of the sgRNA extensively interacts with the positively-charged surface on the back side of the protein (Figure 4-1). In Mol A, the three nucleotides at the 5′ end of the target DNA (3′-ACC-5′, complementary to the PAM) are not visible in the electron density map. Although the two adjacent nucleotides (3′-AC-5′) in Mol B are structurally ordered due to the crystal packing interactions, and are visible in the electron density map, these nucleotides are not recognized by Cas9 (data not shown). These observations suggested that the 3′-NCC-5′ sequence complementary to the 5′-NGG-3′ PAM is not recognized by Cas9, and are consistent with previous biochemical data showing that Cas9-catalyzed DNA cleavage requires the 5′-NGG-3′ PAM on the non-complementary strand but not the 3′-NCC-5′ sequence on the complementary strand (*16*).

Previous studies showed that, although sgRNA with a 48-nt tracrRNA tail [referred to as sgRNA(+48)] is the minimal region for the Cas9-catalyzed DNA cleavage *in vitro* (*16*), sgRNAs with extended tracrRNA tails, sgRNA(+67) and sgRNA(+85), dramatically improved the Cas9 cleavage activity *in vivo* (*40*). The present structure revealed that sgRNA(+48), sgRNA(+67) and sgRNA(+85) contain stem loop 1, stem loops 1–2 and stem loops 1–3, respectively (Figure 4-1A, B). These observations indicated that, whereas stem loop 1 is essential for the formation of the functional Cas9–sgRNA complex, stem loops 2 and 3 further support the stable complex formation and enhance the stability of the sgRNA, thus improving the *in vivo* activity.

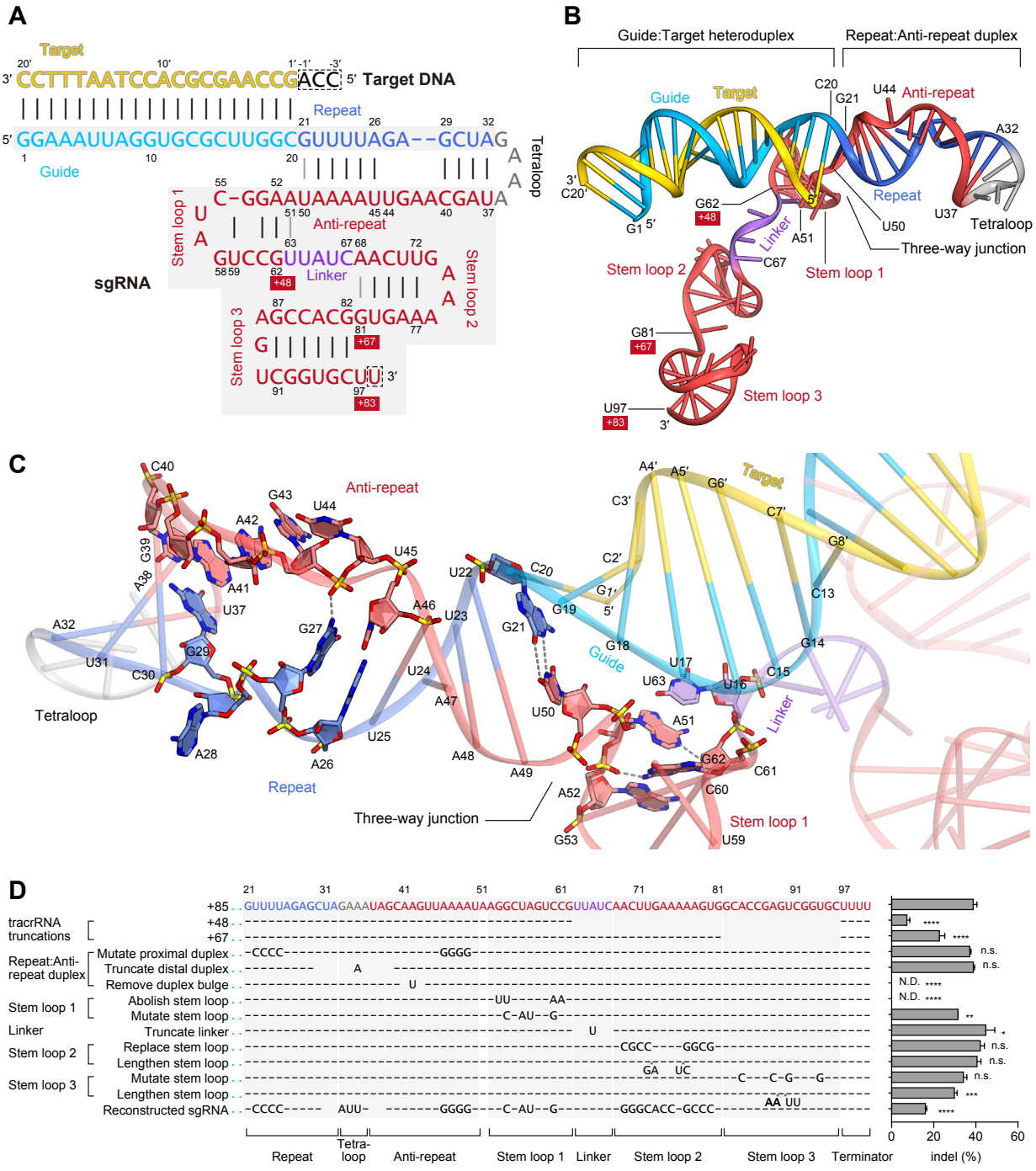


Figure 4-4 sgRNA and target DNA structure

(A) Schematic representation of the sgRNA:target DNA complex. The guide and repeat regions of the crRNA sequence are colored sky blue and blue, respectively. The tracrRNA sequence is colored red, with the linker

region colored violet. The target DNA and the tetraloop are colored yellow and gray, respectively. The numbering of the 3′ tails of the tracrRNA is shown on a red background. Watson-Crick and non-Watson-Crick base pairs are indicated by black and gray lines, respectively. Disordered nucleotides are boxed by dashed lines. (B) Structure of the sgRNA:target DNA complex. (C) Close-up view of the repeat:anti-repeat duplex and the three-way junction. Key interactions are shown with gray dashed lines. (D) Effects of sgRNA mutations on the ability to induce indels. Base changes from the sgRNA(+85) scaffold are shown at the respective positions, with dashes indicating unaltered bases (n = 3, error bars show mean ± S.E.M., *p* values based on unpaired Student's t-test, N.D., not detectable).

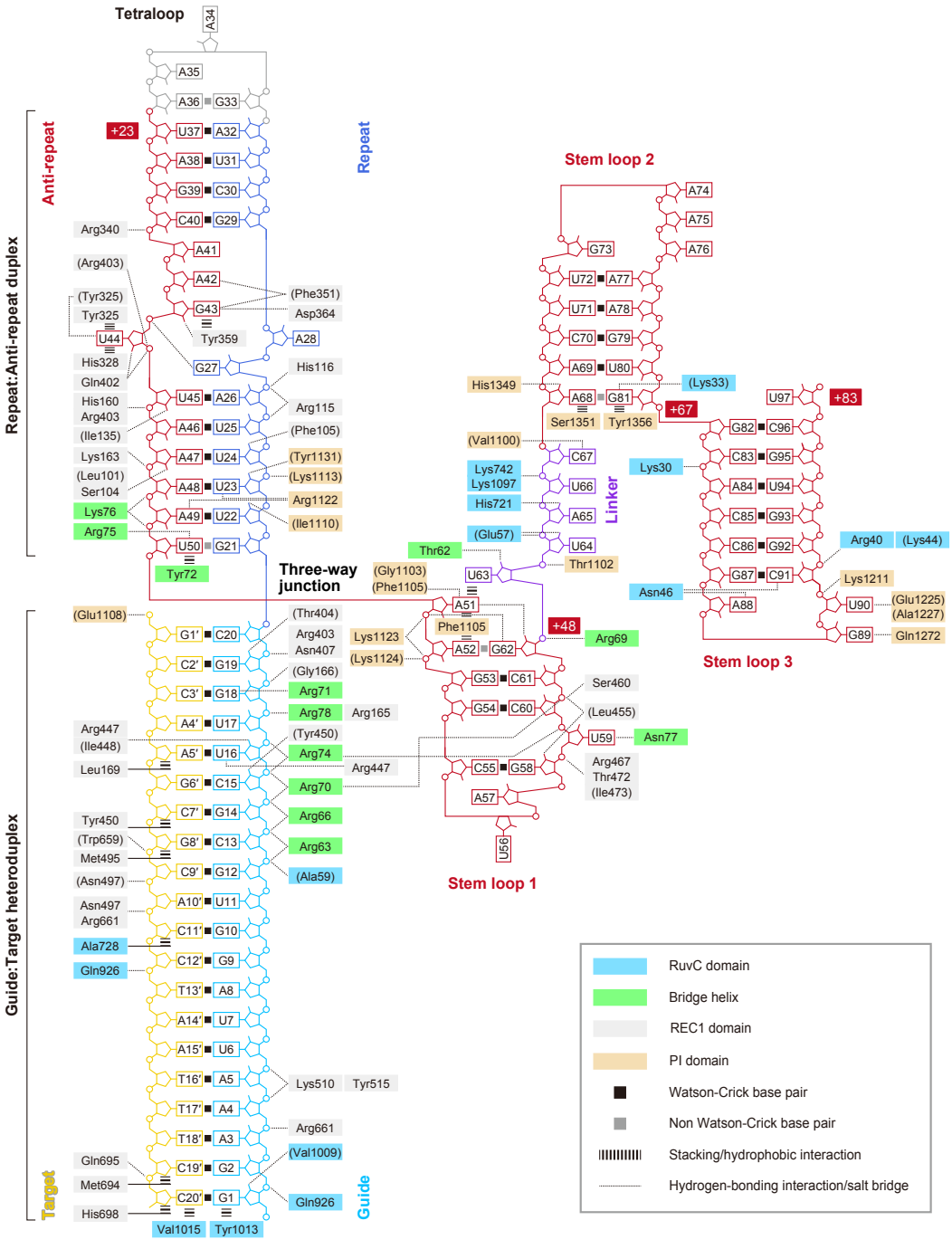


Figure 4-5 Schematic representation of sgRNA:target DNA recognition by Cas9

Residues that interact with the sgRNA:DNA via their main chain are shown in parentheses. Note that water-mediated hydrogen-bonding interactions are not shown, for clarity.

To determine the significance of each sgRNA structural component on the Cas9 function, we tested a number of sgRNAs with mutations in the repeat:anti-repeat duplex, stem loops 1–3, and the linker between stem loops 1 and 2. Our results revealed that, whereas stem loops 2 and 3 as well as the linker region can tolerate a large number of mutations, the repeat:anti-repeat duplex and stem loop 1 are critical for the Cas9 function (Figure 4-1D). Moreover, the sgRNA sequence can tolerate a large number of mutations (Figure 4-1D, reconstructed sgRNA). These results highlight the functional significance of the structure-dependent recognition of the repeat:anti-repeat duplex by Cas9.

The conserved arginine cluster on the Bridge helix is critical for sgRNA:DNA recognition

The sgRNA guide region is primarily recognized by the REC lobe (Figure 4-5). The backbone phosphate groups of the guide region (nucleotides 2, 4–6 and 13–20) interact with the REC1 domain (Arg165, Gly166, Arg403, Asn407, Lys510, Tyr515 and Arg661) and the Bridge helix (Arg63, Arg66, Arg70, Arg71, Arg74 and Arg78) (Figure 4-6A). The 2′-hydroxyl groups of G1, C15, U16 and G19 hydrogen bond with Val1009, Tyr450, Arg447/Ile448 and Thr404, respectively (Figure 4-6A). These observations suggested that the Watson-Crick faces of eight PAM-proximal nucleotides in the Cas9-bound sgRNA are exposed to the solvent, thus serving as a nucleation site for pairing with the complementary strand of the target DNA. This is consistent with previous reports that the 10–12 bp PAM-proximal "seed" region is critical for the Cas9-catalyzed DNA cleavage (*16, 18, 27, 40-42*).

A mutational analysis demonstrated that the R66A, R70A and R74A mutations on the Bridge helix markedly reduced the DNA cleavage activities (Figure 4-6B), highlighting the functional significance of the recognition of the sgRNA "seed" region by the Bridge helix. Although Arg78 and Arg165 also interact with the "seed" region, the R78A and R165A mutants showed only moderately decreased activities (Figure 4-6B). These results are

consistent with the fact that Arg66, Arg70 and Arg74 form multiple salt bridges with the sgRNA backbone, while Arg78 and Arg165 form a single salt bridge with the sgRNA backbone (Figure 4-6A). The cluster of arginine residues on the Bridge helix is highly conserved among the Cas9 proteins in the Type II-A–C systems (Supplementary Figure 13 and Supplementary Figure 14), suggesting that the Bridge helix is a universal structural feature of the Cas9 proteins. This notion is supported by the previous observation that a strictly conserved arginine residue, equivalent to Arg70 of *S. pyogenes* Cas9, is essential for the function of *Francisella novicida* Cas9 in the Type II-B system (*43*). Moreover, the alanine mutations of the repeat:anti-repeat duplex-interacting residues (Arg75 and Lys163) and the stem loop 1-interacting residue (Arg69) resulted in decreased DNA cleavage activity (Figure 4-6B), confirming the functional importance of the recognition of the repeat:anti-repeat duplex and stem loop 1 by Cas9.

The sgRNA guide region is recognized by Cas9 in a sequence-independent manner, except for the U16–Arg447 and G18–Arg71 interactions (Figure 4-5, Figure 4-6A). This base-specific G18–Arg71 interaction may partly explain the observed preference of Cas9 for sgRNAs with guanines in the four PAM-proximal guide region (*25*).

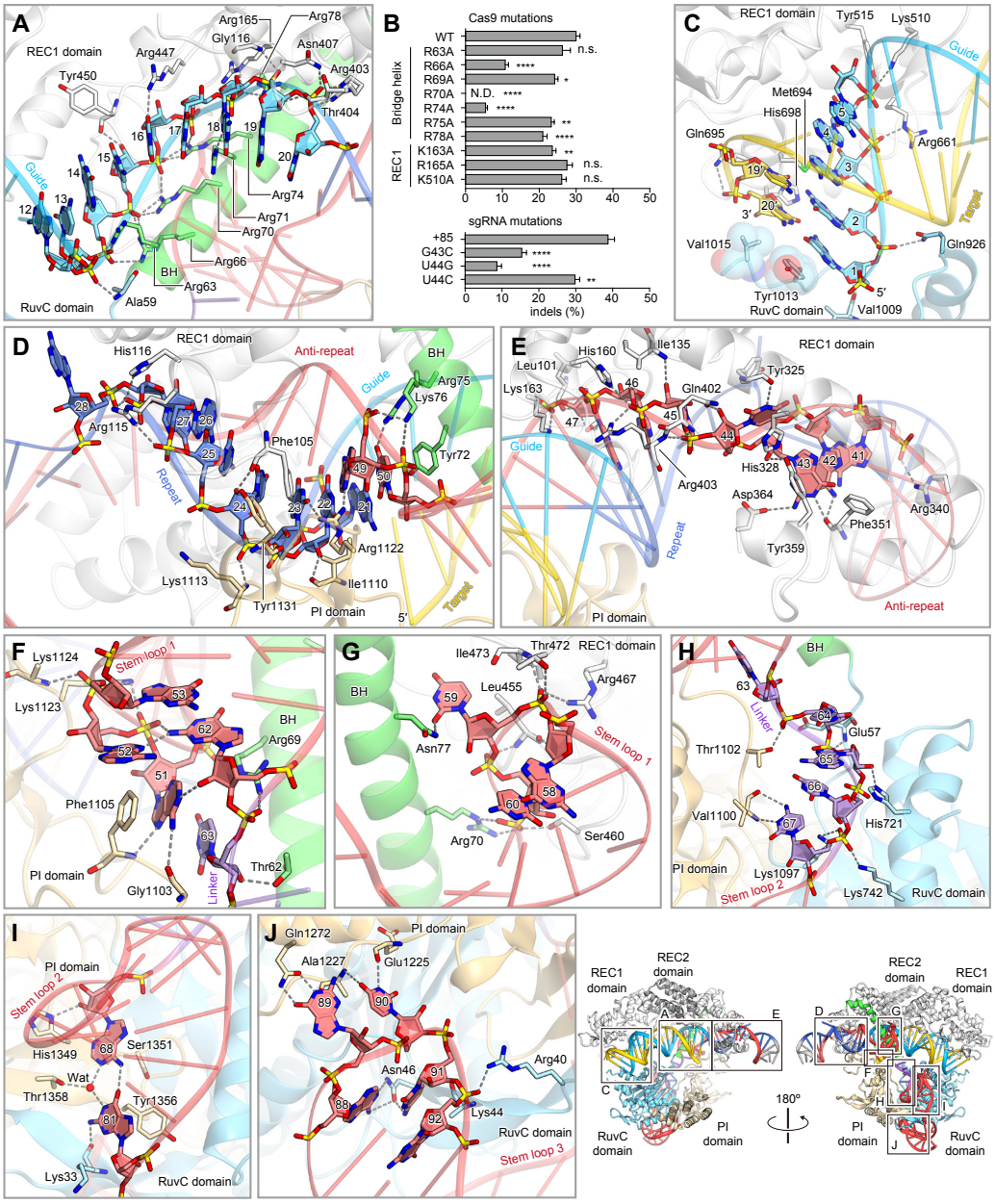


Figure 4-6 sgRNA:target DNA recognition by Cas9

(A and C–J) Recognition of the guide (A), the guide:target heteroduplex (C), the repeat (D), the anti-repeat (E), the three-way junction (F), stem loop 1 (G), the linker (H), stem loop 2 (I) and stem loop 3 (J). Hydrogen bonds and salt bridges are shown as dashed lines. In (A), the target DNA is omitted, for clarity. (B) Effects of Cas9 (top) and sgRNA (bottom) mutations on the ability to induce indels (n = 3, error bars show mean ± S.E.M., *p* values based on unpaired Student's t-test. N.D., not detectable).

The REC1 and RuvC domains facilitate RNA-guided DNA targeting

Cas9 recognizes the 20-bp guide:target heteroduplex in a sequence-independent manner (Figure 4-5). The backbone phosphate groups of the target DNA (nucleotides 1′, 9′–11′, 13′ and 20′) interact with the REC1 (Asn497, Trp659, Arg661 and Gln695), RuvC (Gln926), and PI (Glu1108) domains. The C2′ atoms of the target DNA (nucleotides 5′, 7′, 8′, 11′, 19′ and 20′) form van der Waals interactions with the REC1 domain (Leu169, Tyr450, Met495, Met694 and His698) and the RuvC domain (Ala728) (Figure 4-5). These interactions are likely to contribute toward the ability of Cas9 to discriminate between DNA and RNA targets. The terminal base pair of the guide:target heteroduplex (G1:C20′) is recognized by the RuvC domain via end-capping interactions (Figure 4-6C); the sgRNA G1 and target DNA C20′ nucleobases interact with the Tyr1013 and Val1015 side chains, respectively, whereas the 2′-hydroxyl and phosphate groups of sgRNA G1 interact with Val1009 and Gln926, respectively. These end-capping interactions are consistent with the previous observation that Cas9 recognizes a 17–20-bp guide:target heteroduplex, and that extended guide sequences are degraded in cells and do not contribute to improving sequence specificity (*26*). Taken together, these structural findings explain the RNA-guided DNA targeting mechanism of Cas9.

The repeat:anti-repeat duplex is recognized by the REC and NUC lobes in a sequence-dependent manner

In contrast to the sequence-independent recognition of the sgRNA guide region, sequence-dependent interactions exist between Cas9 and the repeat:anti-repeat duplex (Figure 4-5, Figure 4-6D and 6E). The nucleobases of U23/A49 and A42/G43 hydrogen bond with the side chain of Arg1122 and the main-chain

carbonyl group of Phe351, respectively (Figure 4-6D). The nucleobase of the flipped U44 is sandwiched between Tyr325 and His328, with its N3 atom hydrogen bonded with Tyr325, while the nucleobase of the unpaired G43 stacks with Tyr359 and hydrogen bonds with Asp364 (Figure 4-6E).

The present structure revealed that the repeat:anti-repeat duplex is recognized by the REC lobe, which is divergent in both sequence and length among the Cas9 orthologs within the Type II-A–C systems (Supplementary Figure 13Supplementary Figure 14). This structural finding explains the previous observation that Cas9 and sgRNA are interchangeable only between closely related Type II systems (*34*). The three PAM-distal base pairs (C30:G39–A32:U37) are not recognized by Cas9 and protrude from the complex (Figure 4-5 Schematic representation of sgRNA:target DNA recognition by Cas9), consistent with a proposed model in which the Cas9-bound repeat:anti-repeat duplex is processed by the host RNase III enzyme (*13*).

The nucleobases of G21 and U50, in the G21:U50 wobble pair, stack with the terminal C20:G1′ pair in the guide:target heteroduplex and Tyr72 on the Bridge helix, respectively, with the U50 O4 atom hydrogen bonded with Arg75 (Figure 4-6D). Notably, A51 adopts the *syn*-conformation, and is oriented in the direction opposite to U50 (Figure 4-4C and Figure 4-6F). The nucleobase of A51 is sandwiched between Phe1105 and U63, with its N1, N6 and N7 atoms hydrogen bonded with G62, Gly1103 and Phe1105, respectively (Figure 4-6F). Whereas the repeat:anti-repeat duplexes have diverse sequences and lengths among the Type II-A–C systems, the G:U base pair at the three-way junction is highly conserved among the repeat:anti-repeat duplexes in these three systems (*34*), suggesting that this wobble pairing is a universal structural feature involved in the three-way junction formation.

To verify the importance of the sequence-dependent recognition of the repeat:anti-repeat duplex by Cas9, we evaluated the effects of repeat:anti-repeat mutations on the Cas9-mediated DNA cleavage (Figure 4-6B). The replacement of G43, which forms base-specific hydrogen bonds with Phe351 and Asp364, with cytosine reduced the Cas9 activity by over 3-fold. In addition, the replacement of the flipped U44 with guanine resulted in an over

5-fold drop in the cleavage activity, whereas the replacement of U44 with another pyrimidine base (cytosine) did not significantly affect the cleavage activity (Figure 4-6B). These results confirmed the functional importance of the base-specific recognition of G43 and U44 by Cas9.

Stem loops 1–3 reinforce the interaction between Cas9 and sgRNA

Stem loop 1 is primarily recognized by the REC lobe, together with the PI domain (Figure 4-5). The backbone phosphate groups of stem loop 1 (nucleotides 52, 53 and 59–61) interact with the REC1 domain (Leu455, Ser460, Arg467, Thr472 and Ile473), the PI domain (Lys1123 and Lys1124), and the Bridge helix (Arg70 and Arg74), with the 2′-hydroxyl group of G58 hydrogen bonded with Leu455 (Figure 4-6G). A52 interacts with Phe1105 through a face-to-edge π–π stacking interaction (Figure 4-6F), and the flipped U59 nucleobase hydrogen bonds with Asn77 (Figure 4-6G).

The single-stranded linker and stem loops 2 and 3 are primarily recognized by the NUC lobe (Figure 4-5); in contrast, stem loop 1 and the guide:target/repeat:anti-repeat duplexes are recognized by both the REC and NUC lobes. The backbone phosphate groups of the linker (nucleotides 63–65 and 67) interact with the RuvC domain (Glu57, Lys742 and Lys1097), the PI domain (Thr1102), and the Bridge helix (Arg69), with the 2′-hydroxyl groups of U64 and A65 hydrogen bonded with Glu57 and His721, respectively (Figure 4-6H). The C67 nucleobase forms two hydrogen bonds with Val1100 (Figure 4-6H).

Stem loop 2 is recognized by Cas9 via the interactions between the NUC lobe and the non-Watson-Crick A68:G81 pair, which is formed by direct (between the A68 N6 and G81 O6 atoms) and water-mediated (between the A68 N1 and G81 N1 atoms) hydrogen-bonding interactions (Figure 4-6I). The A68 and G81 nucleobases contact Ser1351 and Tyr1356, respectively, while the A68:G81 pair interacts with Thr1358 via a water-mediated

hydrogen bond (Figure 4-6I). The 2′-hydroxyl group of A68 hydrogen bonds with His1349, while the G81 nucleobase hydrogen bonds with Lys33 (Figure 4-6I).

Stem loop 3 interacts with the NUC lobe more extensively, as compared to stem loop 2 (Figure 4-5). The backbone phosphate groups of C91 and G92 interact with the RuvC domain (Arg40 and Lys44), while the G89 and U90 nucleobases hydrogen bond with Gln1272 and Glu1225/Ala1227, respectively (Figure 4-6J). The A88 and C91 nucleobases are recognized by Asn46 via multiple hydrogen-bonding interactions (Figure 4-6J).

Structural flexibility of Cas9 and sgRNA

Although the HNH domain cleaves the complementary strand of the target DNA at a position three nucleotides upstream of the PAM sequence (*16, 17*), in the present structure, the HNH domain is located away from the scissile phosphate group of the bound complementary strand (Figure 4-7A). A structural comparison of Mol A and Mol B provided mechanistic insights into complementary strand cleavage by the HNH domain. In Mol A, the HNH domain is followed by the α39 helix of the RuvC domain, which is connected to the α40 helix by the α39–α40 linker (residues 919–925) (Figure 4-7A). In Mol A, residues 913–925 form the C-terminal portion of the α39 helix and the α39–α40 linker, while in Mol B, these residues form an extended α-helix, which is directed toward the cleavage site of the complementary strand (Figure 4-7A). These observations suggested that the HNH domain can approach and cleave the target DNA through conformational changes in the segment connecting the HNH and RuvC domains.

The structural comparison further revealed the conformational flexibility between the REC and NUC lobes (Figure 4-7B). As compared to Mol A, Mol B adopts a more open conformation, in which the two lobes are rotated by 15° at a hinge loop between the Bridge helix and strand β5 in the RuvC domain (Figure 4-7B). The bound sgRNA also undergoes an accompanying conformational change at the linker, which interacts with the hinge loop (Figure 4-7C). We also observed the concomitant displacement of the β17–β18 loop of the PI domain,

which interacts with the repeat:anti-repeat duplex and the α2–α3 loop of the REC1 domain (Figure 4-7B). Notably, there is no direct contact between the two lobes in the present structure, except for the interactions between the α2–α3 and β17–β18 loops (Figure 4-7D), suggesting that Cas9 is highly flexible in the absence of the sgRNA. The flexible nature of Cas9 is likely to play a role in the assembly of the Cas9–sgRNA–DNA ternary complex.

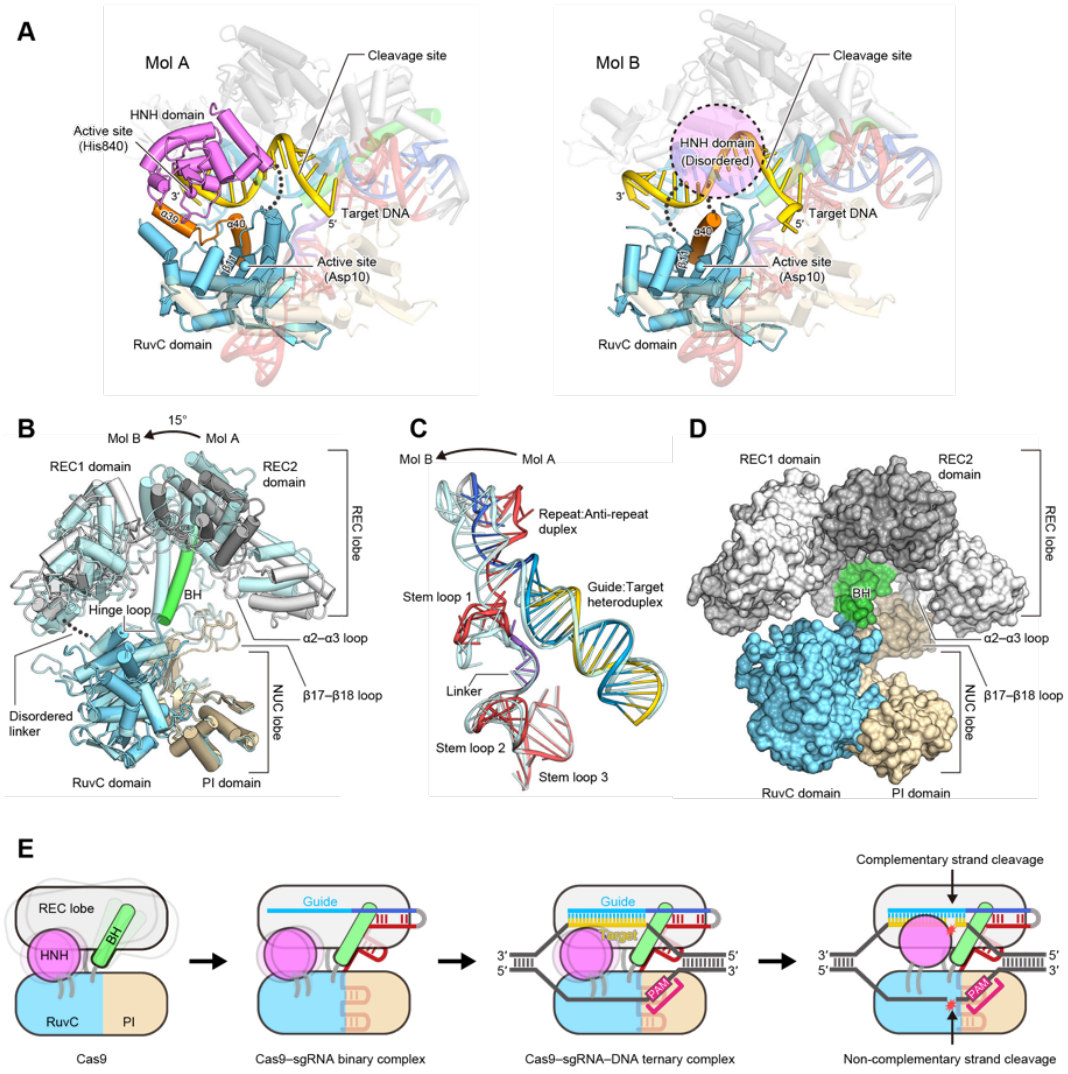


Figure 4-7 Structural flexibility of the complex and a model for RNA-guided DNA cleavage by Cas9

(A) Structural comparison of Mol A and Mol B. In Mol A (left), the disordered linker between the RuvC and HNH domains is indicated by a dotted line. In Mol B (right), the disordered HNH domain is shown as a dashed circle. The flexible connecting segment (a39 and a40) in the RuvC domain is colored orange. (B) Superimposition of the Cas9 proteins in Mol A and Mol B. The two complexes are superimposed based on the core b-sheet of the two RuvC domains. The HNH domain and the bound sgRNA:target DNA complex were omitted, for clarity. (C) Superimposition of the sgRNA:target DNA complex in Mol A and Mol B. After superimposition of the two complexes as in (B), the Cas9 proteins were omitted to show the sgRNA:target DNA complex. (D) Molecular surface of Cas9. The HNH domain and the sgRNA:target DNA complex were omitted, for clarity. (E) Model of RNA-guided DNA cleavage by Cas9.

Discussion

The present structure revealed that the 20-bp heteroduplex, formed by the sgRNA guide region and the complementary strand of the target DNA, is accommodated in the positively-charged groove at the interface between the REC and NUC lobes, with the scissile phosphate group of the target DNA properly positioned for cleavage by the HNH domain. Although the present structure does not contain the non-complementary DNA strand, the position of the bound complementary strand suggested that the scissile phosphate of the non-complementary strand is located in the vicinity of the active site of the RuvC domain, consistent with previous biochemical data (*16, 17*). Furthermore, our structural and functional analyses indicated that the PI domain participates in the PAM recognition.

Based on these observations, we propose a model for the Cas9-catalyzed RNA-guided DNA cleavage (Figure 4-7E). Cas9 recognizes the PAM-proximal guide region and the repeat:anti-repeat duplex of sgRNA, to form the Cas9–sgRNA binary complex. The binary complex subsequently recognizes the DNA sequence complementary to the 20-nt guide region of the bound sgRNA, to form the final Cas9–sgRNA–target DNA ternary complex. Prior to the ternary complex formation, the PI domain recognizes the PAM sequence on the non-complementary strand, thereby triggering the R-loop formation. Upon the assembly of the ternary complex, the mobile HNH domain approaches and cleaves the complementary strand in the guide:target heteroduplex, whereas the RuvC domain cleaves the single-stranded, non-complementary strand. Biochemical studies indicated that PAM recognition by Cas9 is important for both the binding and cleavage of the target DNA, suggesting that the Cas9–sgRNA complex may indeed undergo an inactive-to-active conformational transition upon PAM recognition (*17, 44*). This notion is consistent with the fact that the present structure is likely to represent an inactive state, where the HNH domain is located away from the complementary strand.

The present crystal structure provides a critical step towards understanding the molecular mechanism of RNA-guided DNA targeting by Cas9. Further structural and functional studies with *S. pyogenes* Cas9 or related

orthologs, including the structural elucidation of the Cas9–sgRNA–DNA ternary complex containing the non-complementary strand, will be important for illuminating the mechanisms of PAM recognition, the conformational changes occurring upon PAM recognition, and the mismatch tolerance between the guide:target heteroduplex. However, this study has provided a useful scaffold for the rational engineering of Cas9-based genome modulating technologies. For example, we created an *S. pyogenes* Cas9 truncation mutant (Figure 4-2B) that will facilitate the packaging of Cas9 into size-constrained viral vectors for *in vivo* and therapeutic applications. Moreover, future engineering of the PI domain may allow us to program the PAM specificity, improve the target site recognition fidelity, and increase the versatility of the Cas9 genome engineering platform.

References

1. H. Deveau, J. E. Garneau, S. Moineau, CRISPR/Cas system and its role in phage-bacteria interactions. *Annu Rev Microbiol* **64**, 475 (2010).

2. P. Horvath, R. Barrangou, CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167 (Jan 8, 2010).

3. L. A. Marraffini, E. J. Sontheimer, CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181 (Mar, 2010).

4. M. P. Terns, R. M. Terns, CRISPR-based adaptive immune systems. *Curr Opin Microbiol* **14**, 321 (Jun, 2011).

5. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709 (Mar 23, 2007).

6. S. J. Brouns *et al.*, Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960 (Aug 15, 2008).

7. L. A. Marraffini, E. J. Sontheimer, CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843 (Dec 19, 2008).

8. M. Spilman *et al.*, Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Molecular cell* **52**, 146 (Oct 10, 2013).

9. B. Wiedenheft *et al.*, Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486 (Sep 22, 2011).

10. J. E. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67 (Nov 4, 2010).

11. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39**, 9275 (Nov, 2011).

12. A. Hinnen, J. Hicks, G. Fink, Transformation of yeast. *Proceedings of the National Academy of Sciences of the United States of America* **75**, 1929 (1978).

13. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602 (Mar 31, 2011).

14. R. J. Rothstein, [12] One-step gene disruption in yeast. *Methods in enzymology* **101**, 202 (1983).

15. F. J. Mojica, C. Diez-Villasenor, J. Garcia-Martinez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733 (Mar, 2009).

16. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816 (Aug 17, 2012).

17. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579 (Sep 25, 2012).

18. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

19. P. Mali *et al.*, RNA-guided human genome engineering via Cas9. *Science* **339**, 823 (Feb 15, 2013).

20. S. J. Gratz *et al.*, Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics* **194**, 1029 (Aug, 2013).

21. W. Y. Hwang *et al.*, Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature biotechnology* **31**, 227 (Mar, 2013).

22. H. Wang *et al.*, One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910 (May 9, 2013).

23. H. Yang *et al.*, One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**, 1370 (Sep 12, 2013).

24. O. Shalem *et al.*, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84 (Jan 3, 2014).

25. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80 (Jan 3, 2014).

26. F. A. Ran *et al.*, Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380 (Sep 12, 2013).

27. P. Mali *et al.*, CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology* **31**, 833 (Sep, 2013).

28. L. A. Gilbert *et al.*, CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442 (Jul 18, 2013).

29. S. Konermann *et al.*, Optical control of mammalian endogenous transcription and epigenetic states. *Nature* **500**, 472 (Aug 22, 2013).

30. M. L. Maeder *et al.*, CRISPR RNA-guided activation of endogenous human genes. *Nature methods* **10**, 977 (Oct, 2013).

31. P. Perez-Pinera *et al.*, RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature methods* **10**, 973 (Oct, 2013).

32. L. S. Qi *et al.*, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173 (Feb 28, 2013).

33. L. Holm, P. Rosenstrom, Dali server: conservation mapping in 3D. *Nucleic acids research* **38**, W545 (Jul, 2010).

34. I. Fonfara *et al.*, Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic acids research*, (Nov 22, 2013).

35. M. Ariyoshi *et al.*, Atomic structure of the RuvC resolvase: a holliday junction-specific endonuclease from E. coli. *Cell* **78**, 1063 (Sep 23, 1994).

36. K. M. Gorecka, W. Komorowska, M. Nowotny, Crystal structure of RuvC resolvase in complex with Holliday junction substrate. *Nucleic Acids Res* **41**, 9945 (Nov, 2013).

37. L. Chen, K. Shi, Z. Yin, H. Aihara, Structural asymmetry in the Thermus thermophilus RuvC dimer suggests a basis for sequential strand cleavages during Holliday junction resolution. *Nucleic acids research* **41**, 648 (Jan 7, 2013).

38. C. Biertumpfel, W. Yang, D. Suck, Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature* **449**, 616 (Oct 4, 2007).

39. C. L. Li *et al.*, DNA binding and cleavage by the periplasmic nuclease Vvn: a novel structure with a known active site. *The EMBO journal* **22**, 4014 (Aug 1, 2003).

40. P. D. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827 (Sep, 2013).

41. Y. Fu *et al.*, High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature biotechnology* **31**, 822 (Sep, 2013).

42. V. Pattanayak *et al.*, High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature biotechnology* **31**, 839 (Sep, 2013).

43. T. R. Sampson, S. D. Saroj, A. C. Llewellyn, Y. L. Tzeng, D. S. Weiss, A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**, 254 (May 9, 2013).

44. S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, J. A. Doudna, DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, (2014).

Chapter 5 Efficient *in vivo* genome editing of somatic tissue via Cas9

F. Ann Ran[1,2*], Le Cong[1*], Winston X. Yan[1,3,4*], Bernd Zetsche[1,5], Jonathan S. Gootenberg[1,6], Ophir Shalem[1], Matthias Heidenreich[1], Lukasz Swiech[1], Feng Zhang[1]

[1] Broad Institute of MIT and Harvard,
McGovern Institute for Brain Research,
Department of Brain and Cognitive Sciences,
and Department of Biological Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

[2] Department of Molecular and Cellular Biology
Harvard University
Cambridge, MA 02138, USA

[3] Graduate Program in Biophysics
[4] Harvard-MIT Division of Health Sciences and Technology
[6] Department of Systems Biology
Harvard Medical School
Boston, MA 02115, USA

[5] Department of Developmental Pathology, Institute of Pathology,
University of Bonn, Bonn, Germany

* These authors contributed equally to this work.

Acknowledgements

We thank Abigail Scherer-Hoock and the MIT Division of Comparative Medicine for assistance with animal experiments, Chie-Yu Lin for experimental assistance, and the entire Zhang lab for support and advice. W.X.Y. is supported by award Number T32GM007753 from the National Institute of General Medical Sciences. J.S.G. is supported by the National Institutes of Health Grant GM080177. This work is supported by an NIH Director's Pioneer Award (5DP1-MH100706), a NIH Transformative R01 grant (5R01-DK097768), the Keck, McKnight, Damon Runyon, Searle Scholars, Klingenstein, Merkin, Vallee, CL, and Simons Foundations, Bob Metcalfe, and Jane Pauley.

Author Contributions

F.A.R., L.C., W.X.Y., B.Z., J.S.G., and F.Z. designed and performed the experiments. F.A.R., L.C., W.X.Y., and F.Z. analyzed the data. J.S.G. contributed computational analysis of PAM sites and off-target analysis. O.S., F.A.R., and F.Z. performed metagenomic analysis of Cas9, crRNA, and tracrRNA orthologs. F.A.R., W.X.Y., L.C., and F.Z. wrote the manuscript with help from all authors. Specifically, I contributed the crRNA/tracrRNA predictions for the Cas9 orthologs, designed and carried out the *in vitro* PAM screen, biochemical and mammalian cell culture validation of orthologs, specificity analysis, and experiments in mammalian cells.

Introduction

The CRISPR (clustered regularly interspaced short palindromic repeats)-Cas system is a RNA-guided endonuclease system from bacteria and archaea that provides adaptive immunity against exogenous nucleic acids(*1-6*). Of the three CRISPR-Cas classes, the Type II system has to date attracted the most interest as a genome engineering platform because of its relatively simple and well-characterized mechanism – a single endonuclease (Cas9) and two small RNAs, the CRISPR RNA (crRNA) that contains the DNA-targeting guide sequence (spacer) and the auxiliary trans-activating crRNA (tracrRNA)(*7*), mediate cleavage of the target DNA (protospacer)(*7-9*); this dual RNA complex has been further engineered into a chimeric single-guide RNA (sgRNA)(*10-12*). An additional requirement critical to Cas9 activity is the presence of a protospacer adjacent motif (PAM) in the target DNA, which differs among the CRISPR-Cas systems(*5, 13*).

The ability to harness Cas9 for broad applications *in vivo* in somatic tissue, while obviating the need for embryonic manipulation, would prove enormously useful for accelerating basic research and enabling clinical applications(*14, 15*). One major challenge is the delivery of the Cas9 genome editing system to animals. Adeno-associated virus (AAV) vectors are attractive candidates for efficient gene delivery *in vivo* because of their low immunogenic potential, reduced oncogenic risk from host-genome integration(*16*), and well-characterized serotype specificity(*17-20*). However, the limited cargo size of ~4.5kb for optimal transgene delivery renders the packaging of SpCas9 (~4.2kb) and appropriate control elements (promoter, polyA signal) difficult. While several smaller Cas9 orthologs have been used for mammalian genome editing(*10, 21*), they are nonetheless relatively limited in availability of targeting sequences due to the requirement for lengthier and more specific PAMs, and cannot match SpCas9 in cleavage efficiency. This highlights the potential as well as the need to further explore the ecological diversity of Type II CRISPR systems for additional suitable Cas9s.

Here, we have identified six small Cas9 orthologs and their corresponding protospacer adjacent motifs (PAM), which could be optimized for mammalian genome editing. In particular, we show that Cas9 from *Staphylococcus*

aureus (SaCas9), which is 23% smaller than SpCas9, can edit the mammalian genome with high efficiency on par with SpCas9, and be packaged along with its single-guide RNA (sgRNA) into adeno-associated virus (AAV) as a single vector for delivery into adult mice. We demonstrate targeting of the mouse liver and observed 30% gene modification *in vivo* within 3 weeks of injection.

Metagenomic search for small Cas9 orthologs

To identify a diverse set of small Cas9 proteins, we selected six representative Cas9 orthologs from over 800 known Cas9s from GenBank and optimized their sequences for mammalian expression (Figure 5-1a). These Cas9s belong to the Type IIA and IIC subfamilies(*22, 23*). Using the characteristic direct repeat motifs found within the CRISPR array(*2, 3, 24*), we searched a 2-kb window flanking the CRISPR locus for potential tracrRNAs that contained strong sequence homology to the repeats, at least two additional predicted stemloops, and a Rho-independent transcriptional termination signal within 150-nt. From these we constructed sgRNA scaffolds for each ortholog (Supplementary Figure 16 and Supplementary Table 6). Since the full 3' end of tracrRNA improves sgRNA abundance in cells(25) and mediates interaction with Cas9(*26*), we included the full tracrRNA 3' end for each ortholog. We then cleaved a library of plasmids containing a fixed-sequence target followed by a randomized 7-mer as PAM (5'-NNNNNNN) in an *in vitro* cell lysate assay, and identified the putative PAMs by sequencing the targets that were successfully cleaved (Figure 5-1b, c). We observed that similar to SpCas9, the Cas9 orthologs cleaved targets 3 bp upstream of PAM (Supplementary Figure 17). To validate the consensus PAMs from the library, we subsequently cleaved a DNA template bearing the putative PAMs in a biochemical lysate reaction and showed that the sgRNA designs, in combination with the Cas9 orthologs, can indeed target sites bearing appropriate consensus PAMs, albeit with differing efficiencies (Figure 5-1d and Supplementary Table 7).

.

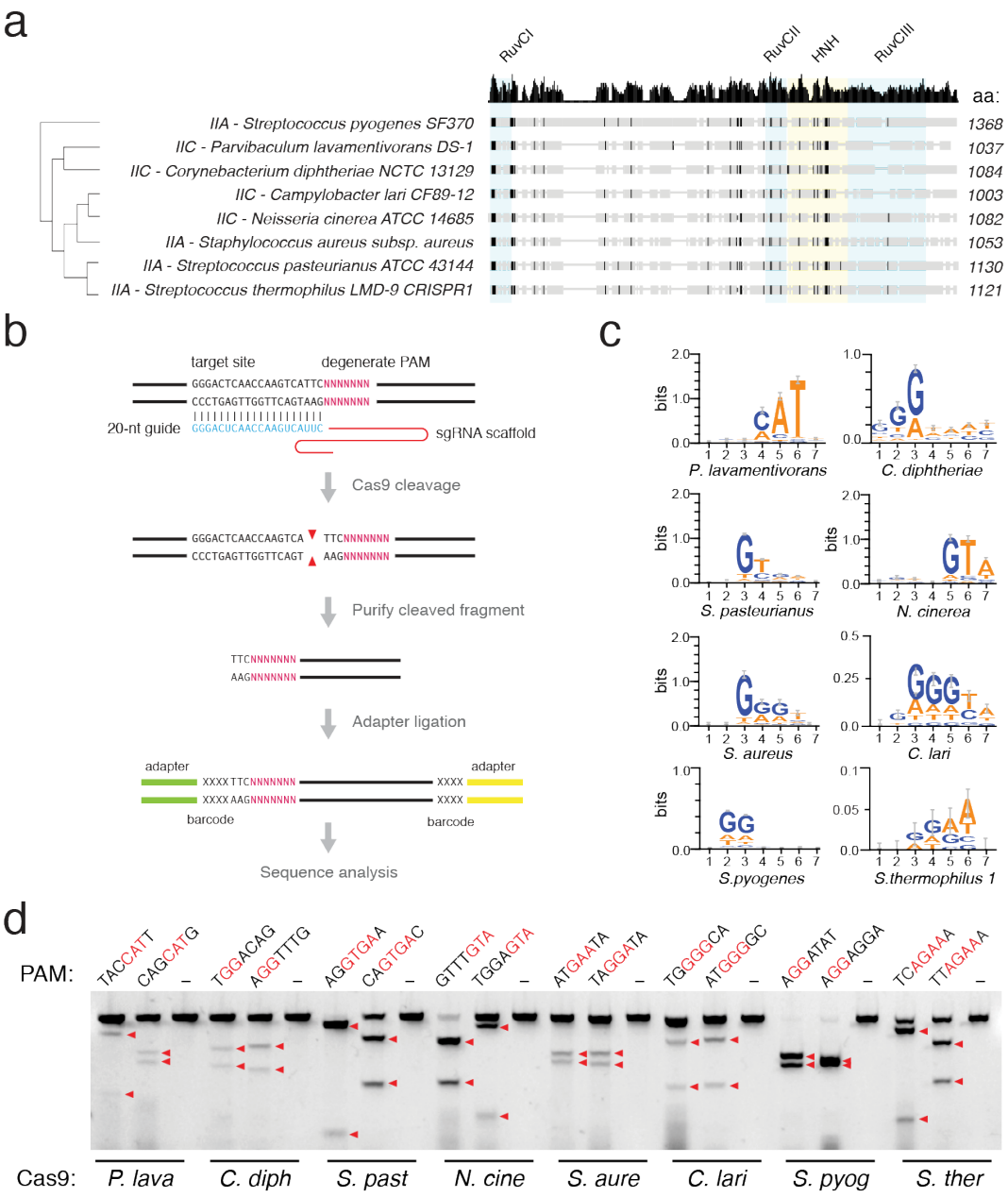


Figure 5-1 Biochemical screen for small Cas9 orthologs

(A) Phylogenetic tree of Cas9 orthologs, with subfamily and sizes (amino acids) indicated. Conserved nuclease domains are highlighted in colored boxes, black residues represent conserved sequences. (B) Schematic illustrating *in vitro* cleavage-based method used to identify protospacer adjacent motifs (PAMs). (C) Consensus PAMs for eight Cas9 orthologs from sequencing of cleaved fragments. (D) Biochemical cleavage reaction using orthologs and sgRNAs targeting different loci bearing the putative PAMs (shown in red). Red triangles indicate cleavage fragments.

Characterization of SaCas9 in vitro

Having validated the activity of Cas9 orthologs using cell lysates, we sought to test their ability to induce double stranded breaks in mammalian cells. We co-transfected in human embryonic kidney (HEK 293FT) cells the Cas9 orthologs and their respective sgRNAs targeting endogenous human genomic loci with the appropriate PAMs. However, of the six Cas9 orthologs tested, only the Cas9 from *Staphylococcus aureus* (referred to as SaCas9) reproducibly yielded indels by SURVEYOR assay (Supplementary Figure 18 and Supplementary Table 8). Thus, we focused on optimizing SaCas9 and sgRNA for application in *in vivo* mammalian genome editing.

Although many Type II CRISPR systems share a common feature of having ~36-bp direct repeats and ~30-bp spacers(*22*), previous studies have reported different lengths for spacer as well as direct repeat sequences in the mature crRNA among different systems(*7, 21, 27*). We therefore sought to test the optimal lengths of these two parameters for the SaCas9 sgRNA (Figure 5-2a). We found that while a range of spacer or guide length is tolerated for SaCas9, there is a marked decrease in cleavage efficiency when it is 18-nt or below (Figure 5-2b), in contrast to SpCas9 where shorter sgRNA lengths can be used(*28*). Similarly, a range of lengths for direct repeat:tracrRNA antirepeat duplex is tolerated (Figure 5-2c). Based on these results, we chose the shorter 20-nt guide, 14-bp repeat:antirepeat duplex sgRNA architecture for downstream applications.

Since there might be potential differences between the cell lysate and the endogenous mammalian nuclei environment that may affect DNA cleavage specificity, we wanted to verify whether the *in vitro* 5'-NNGRR(T) consensus PAM held for SaCas9 cleavage in mammalian cells. From SURVEYOR analysis of endogenous genome cleavage based on 116 distinct genomic target sites, we determined that SaCas9 could efficiently cleave genomic targets with a 5'-NNGRR PAM, with no requirement for the T in the 6th position (Figure 5-2d and Supplementary Table 9). On average, the 5'-GRR motif occurs in the human genome every 7.6-bp, allowing the SaCas9 to have a wide range of available targets (Supplementary Figure 19).

Among the Cas9 orthologs used for mammalian genome editing, SpCas9 remains the best characterized in targeting specificity, with consistently high editing efficiency across multiple cell types and species. For three targets in mouse hematoma (Hepa1-6) cells, the editing efficiency of SaCas9 performed comparably with that of SpCas9 (Figure 5-2e). Furthermore, we assayed genomic off-target indel mutations at highly similar genomic sequences(25) for both SaCas9 and SpCas9, targeting a common locus bearing an overlapping 5'-NGGRR PAM. At 31 genome-wide loci with sequence similarity to intended target, SaCas9 cleaved off-target sites with comparable activity as SpCas9 (Figure 5-2f and Supplementary Table 10).

Adeno-associated virus delivery of SaCas9 in vivo

Having established and validated the optimal sgRNA architecture for SaCas9 in mammalian cells, we sought to incorporate SaCas9 into AAV vector for *in vivo* use. In AAV, the small size of SaCas9 (3.2kb) leaves sufficient room for promoters of up to 600-bp in a dual-cassette design co-expressing SaCas9 and U6-driven sgRNA (Figure 5-3a). The ability to apply Cas9 protein to modify endogenous loci in somatic tissues or adult animals enables rapid testing of gene function in the relevant tissue type and therapeutic applications for gene correction. Of the organs targetable by AAV, the liver is particularly attractive for demonstrating the feasibility and therapeutic potential of CRISPR-Cas mediated *in vivo* genome engineering because of its accessibility by intravascular delivery and its central role in many metabolic pathways important for human disease(29). We chose to target the mouse locus encoding proprotein convertase subtilisin/kexin type 9 (*Pcsk9*), an enzyme that is predominantly expressed in the liver and involved in cholesterol homeostasis, whose reduction has shown promise in lowering the risk of cardiovascular disease(*30, 31*).

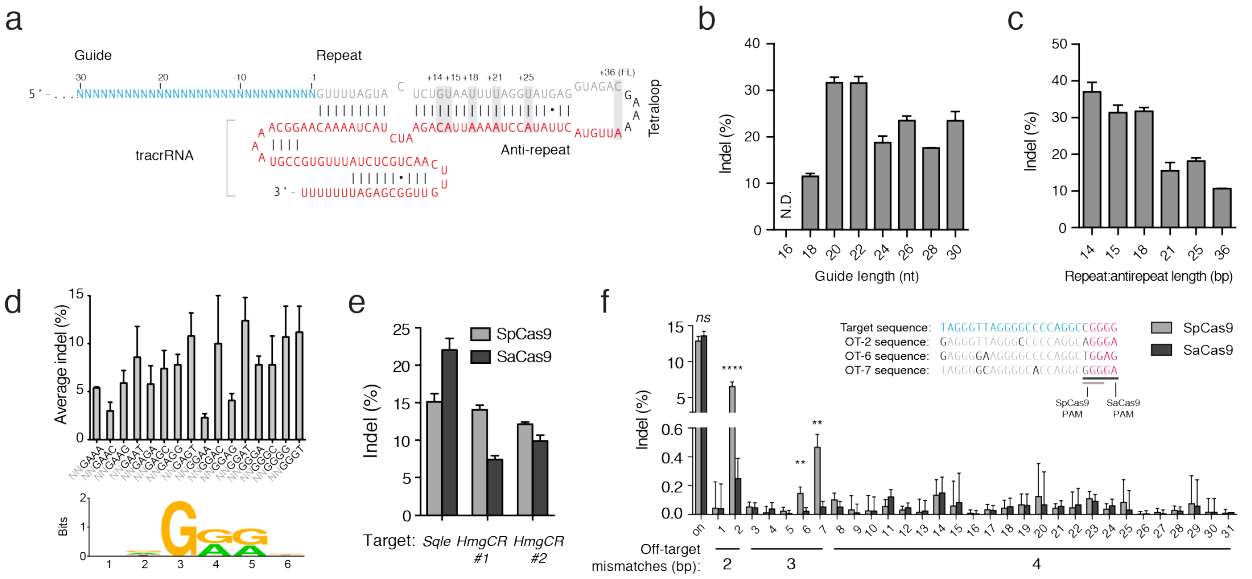


Figure 5-2 in vitro characterization of Staphylococcus aureus Cas9

(A) Schematic showing the structure of *S. aureus* sgRNA. Indels vary depending on (B) length of guide sequence or (C) repeat:anti-repeat duplex. (D) Consensus PAM for SaCas9 in HEK 293FT cells. Pooled indel values for all putative PAM 4-base combinations (top, $n \geq 3$) and overall sequence logo ($n = 116$, bottom) are shown. SpCas9 and SaCas9 cleavage efficiency comparison for (E) genomic target sites and (F) genome-wide off-target loci (error bars indicate Wilson intervals). Off-target (OT) sequences with significant indels are highlighted above graph. $n = 3$, error bars S.E.M unless otherwise noted; N.D. not detectable.

Using AAV2/8, a highly efficient hepatotropic AAV serotype, we delivered via tail-vein injection 8×10^{10} viral particles using single-vector design containing a cytomegalovirus (CMV) promoter-driven SaCas9 and a U6 promoter-driven sgRNA targeting *Pcsk9* (Figure 5-3a, b). The percentage indel formation increased from approximately 5% at 1 week to 28% at 11 weeks, demonstrating the *in vivo* editing capabilities of SaCas9 and the single-vector design (Figure 5-3c). To further increase the efficiency of genome modification, we screened additional guides targeting *Pcsk9* in Hepa1-6 cells (Supplementary Figure 20) and used a liver-specific thyroid-binding globulin (TBG) promoter to provide greater hepatocyte specificity and expression(*32, 33*). After intravascular delivery of 2×10^{11} viral particles, we observed indel formation in the liver ranging from 11% at 1-week post injection to approximately 30% at 3 weeks (Figure 5-3c-e). The *Pcsk9* gene modification level remained consistent across samples from multiple locations within the liver, suggesting that the delivery was uniform throughout the target organ (Figure 5-3d). All mice survived the AAV injection and did not exhibit any signs of physical distress for the entire duration of the experiment.

Discussion

The small size and efficiency of the novel Cas9 ortholog from *S. aureus* paves the way for rapid and versatile *in vivo* editing while maintaining target specificity through promoter and AAV serotype selection. Furthermore, the method of PAM identification described here presents a generalizable approach to PAM identification amongst all Type II CRISPR systems. While certain Cas9 orthologs are more readily adapted for mammalian genome editing than others, SaCas9 cleaves endogenous targets in cells with robust efficiencies similar to those of SpCas9 and additionally exhibits a similar degree of specificity. However, additional studies are necessary to fully characterize the specificity of SaCas9 as well as the effects of prolonged Cas9 *in vivo* expression.

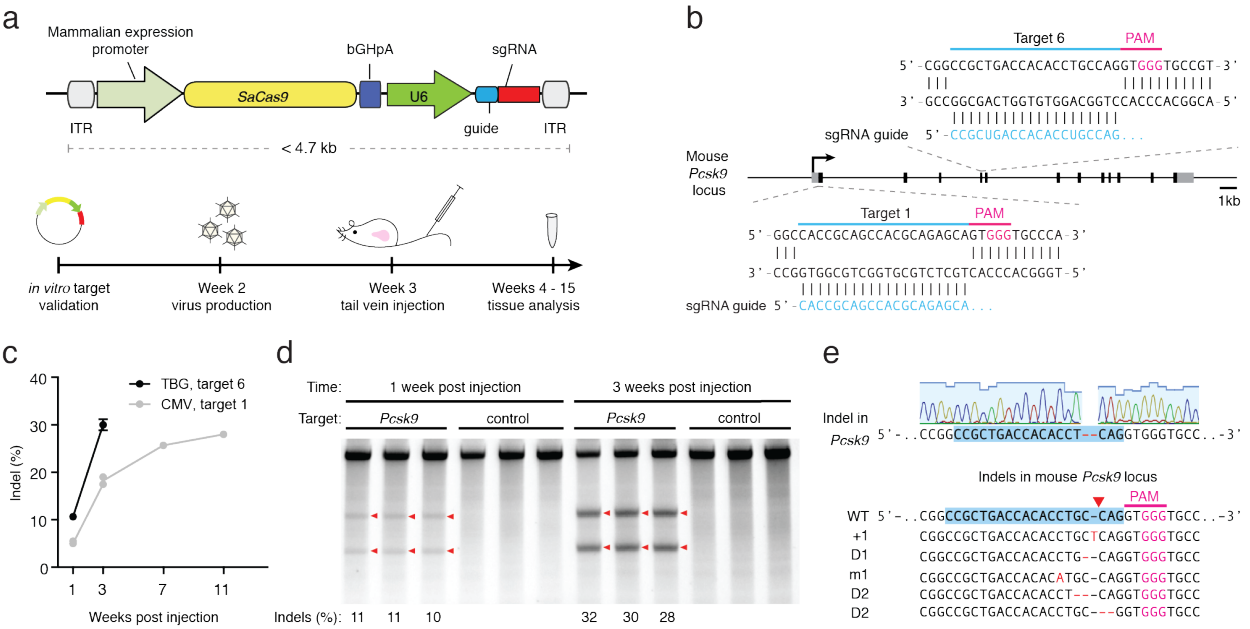


Figure 5-3 AAV delivery of S. aureus Cas9 into live animals

(A) Schematics illustrating AAV single-vector system (top) and experimental timeline (bottom). (B) Mouse *Pcsk9* locus showing SaCas9 target locations. Guide sequences are highlighted in blue with PAM in magenta. (C) Time course of liver tissue indel formation at targets 1 and 6 post injection of AAV2/8 particles (up to 2 animals each; error bars represent liver tissue pieces). (D) Indel formation at target 6 at 1 and 3 weeks post-injection. Each lane represents a piece of liver tissue. Red triangles indicate cleavage fragments. (E) Representative chromatogram and indels generated by SaCas9 *in vivo*.

While the AAV-delivery of the Cas9 system is a promising step towards gene therapy applications, the more immediate impact lies in the efficient interrogation of genetic contributions to both normal biology and disease in animals beyond cell lines and transgenic models. Such somatic or postnatal genetic manipulation allows unprecedented spatial and temporal control of targeted gene modifications that may be developmentally important or inadequately controlled by conditional expression systems, as well as the ability to simulate a gradual accumulation of genetic mutations that could better model the natural progression of certain pathogenic processes. Lastly, viral vector mediated gene modification allows for significantly higher throughput of studying genetic variants of disease than transgenic animal generation, particularly in organisms with lengthy gestational and developmental periods. The *in vivo* opportunities made possible by the AAV delivery of the *S. aureus* Cas9 described here represents another piece of the continually expanding Cas9 genome engineering toolbox that promises to allow rapid advances across basic science, medical, and biotechnology applications.

References

1. L. Marraffini, E. Sontheimer, CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (New York, N.Y.)* **322**, 1843 (2008).

2. R. Barrangou *et al.*, CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)* **315**, 1709 (2007).

3. S. J. Brouns *et al.*, Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science (New York, N.Y.)* **321**, 960 (2008).

4. C. Hale *et al.*, RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945 (2009).

5. F. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology (Reading, England)* **155**, 733 (2009).

6. R. Jansen, J. Embden, W. Gaastra, L. Schouls, Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology* **43**, 1565 (2002).

7. E. Deltcheva *et al.*, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602 (2011).

8. J. Garneau *et al.*, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67 (2010).

9. R. Sapranauskas *et al.*, The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic acids research* **39**, 9275 (2011).

10. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

11. P. Mali *et al.*, RNA-guided human genome engineering via Cas9. *Science* **339**, 823 (Feb 15, 2013).

12. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816 (2012).

13. A. Bolotin, B. Quinquis, A. Sorokin, S. Ehrlich, Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology (Reading, England)* **151**, 2551 (2005).

14. V. Bedell *et al.*, In vivo genome editing using a high-efficiency TALEN system. *Nature* **491**, 114 (2012).

15. H. Li *et al.*, In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature* **475**, 217 (Jul 14, 2011).

16. A. Vasileva, R. Jessberger, Precise hit: adeno-associated virus in gene targeting. *Nat Rev Microbiol* **3**, 837 (Nov, 2005).

17. F. Mingozzi, K. A. High, Therapeutic in vivo gene transfer for genetic disease using AAV: progress and challenges. *Nature reviews. Genetics* **12**, 341 (May, 2011).

18. G. Gao, L. H. Vandenberghe, J. M. Wilson, New recombinant serotypes of AAV vectors. *Current gene therapy* **5**, 285 (Jun, 2005).

19. M. A. Kay, State-of-the-art gene-based therapies: the road ahead. *Nature reviews. Genetics* **12**, 316 (May, 2011).

20. C. Zincarelli, S. Soltys, G. Rengo, J. E. Rabinowitz, Analysis of AAV serotypes 1-9 mediated gene expression and tropism in mice after systemic injection. *Mol Ther* **16**, 1073 (Jun, 2008).

21. Z. Hou *et al.*, Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 15644 (2013).

22. I. Fonfara *et al.*, Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic acids research* **42**, 2577 (2014).

23. E. V. Koonin, K. S. Makarova, CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol* **10**, 679 (May, 2013).

24. L. Marraffini, E. Sontheimer, CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews. Genetics* **11**, 181 (2010).

25. P. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827 (2013).

26. H. Nishimasu *et al.*, Crystal structure of cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935 (2014).

27. K. Chylinski, A. Le Rhun, E. Charpentier, The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA biology* **10**, 726 (2013).

28. Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio, J. K. Joung, Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature biotechnology*, (Jan 26, 2014).

29. T. H. Nguyen, N. Ferry, Liver gene therapy: advances and hurdles. *Gene Ther* **11 Suppl 1**, S76 (Oct, 2004).

30. J. C. Cohen, E. Boerwinkle, T. H. Mosley, Jr., H. H. Hobbs, Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *The New England journal of medicine* **354**, 1264 (Mar 23, 2006).

31. M. Frank-Kamenetsky *et al.*, Therapeutic RNAi targeting PCSK9 acutely lowers plasma cholesterol in rodents and LDL cholesterol in nonhuman primates. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 11915 (Aug 19, 2008).

32. Y. Hayashi *et al.*, Human thyroxine-binding globulin gene: complete sequence and transcriptional regulation. *Molecular endocrinology* **7**, 1049 (Aug, 1993).

33. R. J. Chandler *et al.*, Liver-directed adeno-associated virus serotype 8 gene transfer rescues a lethal murine model of citrullinemia type 1. *Gene Ther* **20**, 1188 (Dec, 2013).

Chapter 6 Perspectives and future directions

The processes of life – an extraordinarily intricate balance of development and differentiation, adaptation and homeostasis, replication and senescence – are governed by an equally complex network of genetic interactions coupled with environmental influences. The quest to understand the causal relationships between genotype and phenotype in both normalcy and aberrancy has continually driven the development of technologies that have allowed us to ask ever more fine-mapped questions about the genome. In the past several decades, two avenues of advancements have had enormous impacts on our ability to study molecular processes at a more refined level. Accompanying the exponential growth of computing capacity, revolutions in sequencing technology, particularly the development of deep-sequencing platforms, have enabled the whole-genome mapping of an ever-increasing number of species far beyond the ensemble of traditional model organisms. As the depth of coverage increases, we are also able to attain a more refined understanding of the nucleotide differences among members within a species, tissue types within an individual, and health and disease within a cell type. This wealth of information is only beginning to be analyzed by groups worldwide, including the tremendously exciting and ambitious effort to annotate the human genome by the ENCODE consortium. However, our ability to functionally parse the sequences relies not only on making associations with existing variations, but observing phenotypic changes arising from targeted, deliberate alterations. Concurrent with the development of sequencing technologies, the molecular biology tools and techniques we use for interrogating gene function have been greatly expanded and refined. This thesis is then an account of just a part of this continued evolution, and what follows is an overview of the current state of the field, the continued effort at improving the technology by myself, my colleagues, and others, and some of my personal thoughts on exciting possibilities these technologies have opened up as well as some of the challenges facing genome editing.

A new approach to genome engineering

In the last year and half, we and others in the field have spent much of our effort focusing on the development, characterization, and validation of CRISPR-Cas as a viable and robust genome engineering tool. The qualities we

looked for in a truly useful genome editing technology were efficiency, versatility, and specificity. Though both ZFNs and TALENs were efficient and potentially quite specific at mediating DNA DSB, their major weaknesses came from their relatively high barrier to entry: a considerable amount of design and cloning was involved in the construction and testing of each new pair. At the time, even the most accessible TALEN technology cost about $5000 per target site per our own calculations.

The groundbreaking *Science* study from Emmanuelle Charpentier and Jennifer Doudna on the *in vitro* characterization of the CRISPR nuclease system on the other hand had raised a tantalizing possibility for more generalized applications (*1*). The idea of a mechanism similar to RNAi for gene editing, dubbed "DNAi" by Virginijus Siksnys, one of the pioneers of CRISPR biology (*2*), had drawn enormous interest in the field, promising a new tool that could potentially overcome many of the challenges facing earlier generation designer nucleases in customizability, affordability, and simplicity of design. CRISPR technology would make it possible to target specific sites in the mammalian genome with merely a pair of DNA oligos designed with minimal effort that cost less than $30.

In reporting the first development and demonstration of CRISPR for mammalian genome engineering, we hoped that this technology would further facilitate the "open-source genome engineering" begun by its TALEN and OPEN-ZFN predecessors, allowing any lab to carry out studies functionally validating a gene of interest, generate new cell lines for disease and therapeutic modeling, and manipulate model organisms at will. Given its ease of application, CRISPR has indeed been widely adopted and used with stunning rapidity in a host of model systems and species, proving itself quite efficient and versatile and perhaps even more so than ZFNs and TALENs.

The specificity of Cas9, however, remains one of the most immediate and important questions that has yet to be fully addressed. In reports from bacterial CRISPR manipulation (*1, 3-5*) as well as our initial mammalian experiments, we proposed that a PAM-proximal "seed region" of 8-10 base pairs between the guide sgRNA and target DNA was most important to the specificity of Cas9 cleavage: mismatches within this region would disrupt

Cas9 cleavage. However, more comprehensive analyses reveal that while specificity is weighted toward the "seed region," it is determined by a combination of factors including the number, identity, spacing, and location of mismatches (*6, 7*), and there doesn't seem to exist a simple, generalizable design rule. Recently, Joung and colleagues reported potentially reduced off-target effects using "tru-sgRNAs" or sgRNAs with truncated guide region (e.g. 18-bp instead of 20-bp of total recognition). It is important to stress, however, that most of these studies were performed with predicted off-targets based on similarity to the 20-bp target sequence, and an unbiased, genome-wide analysis – especially in the case of tru-sgRNAs – has yet to be undertaken. Given the breadth of potential off-target sites and depth of coverage needed to differentiate noise from signal, such an undertaking based on indel analysis would remain technically exceedingly challenging. For this purpose, perhaps alternative approaches such as the recently reported BLESS (direct *in situ* breaks labeling, enrichment on streptavidin and next-generation sequencing) technique might be explored (*8*).

Better characterization of Cas9 specificity and toxicity will be especially crucial for certain applications. One obvious such example is the field of gene therapy (discussed later), where unintended modifications could lead to disastrous consequences. In addition, effects of long term *in vivo* expression of Cas9 remain unknown: could the nuclease co-opt endogenous microRNAs, leading to unpredictable activity? Others, such as the use of dead Cas9 as a generic DNA-binding module for transcriptional regulation, cannot rely on an obligate dimerization (double-nicking) strategy to reduce off-target activity, and there the specificity of Cas9 or sgRNA must be better defined. To this end, the structural analysis of Cas9:sgRNA:DNA complex might shed light on the possibility of protein evolution or rational engineering; for instance, the arginine-rich "bridge helix" in the recognition domain of SpCas9 makes extensive non-sequence-specific contacts with the sgRNA backbone, and perhaps an analysis of those interactions may contribute to understanding Cas9 specificity.

For on-target modification, one of the only known requirements for Cas9 activity is the presence of the PAM. For SpCas9, the 5'-NGG-3' motif may occur frequently enough in the genome for many general applications, but nevertheless remains a constraint in some cases, *e.g.* HDR where frequency falls rapidly away from site of DSB and

genomic microdeletions where precise boundaries of resection are desired. For other orthologs such as St1Cas9 (5'-NNAGAAW-3') or NmCas9 (5'-NNNNGATT-3'), the PAM constraint becomes much more restrictive. The possibility of altering or removing the PAM requirement altogether would make Cas9 a truly universal programmable nuclease. Again, the structure study of SpCas9 and those of additional orthologs in the future might provide some insight on the potential importance of the C'-terminal domain in interactions with the PAM-bearing DNA strand. It remains to be seen to what degree it may be manipulated and Cas9 chimeras generated. Here, the mining for metagenomic diversity of Cas9 for PAM diversity as well as enhanced enzyme specificity and efficiency will prove valuable.

While Cas9 almost invariably cleaves any target with a requisite PAM, we and others have observed some differences in cleavage efficiency among targets and that a small number of 20-bp sites do not get cleaved by Cas9. Though we have observed from a limited dataset both *in vitro* and in cells that Cas9 does not appear to be sensitive to DNA methylation, other factors that influence Cas9 targeting efficiency remain largely unknown, and could arise from properties intrinsic to sgRNA, target DNA, or a combination thereof. Systematic studies examining parameters such a guide sequence thermodynamics and structure, genomic context, and chromatin state will likely be enlightening.

It's also important to note that although Cas9 has enabled the efficient and easy induction of targeted DSBs, we still have little control over the downstream repair events. Our ability to make ultraprecise genome modifications with pre-determined, predictable outcomes will likely depend on new methods to predispose the cell to undergo HDR rather than NHEJ. Although the use of nickase is a step towards this, the efficiency of HDR in most contexts, particularly in human and mouse ES cell lines, remains quite low. Meanwhile, studies overexpressing HDR proteins to stimulate recombination have found mixed results, with discrepancies potentially arising from differences in the model systems, the species from which the transgenes are derived, the dosage of overexpression, and the nature of the DNA damage (*9-11*). Thus, the development of methods for more precise temporal and

spatial modulation of the cellular NHEJ and HDR machinery, as well as better delivery and optimization of donor templates, may pave the way for higher efficiency of precision gene modification.

The application of CRISPR-Cas for genome engineering is still nascent; nevertheless, the field has in a relatively short time developed a general understanding of Cas9 function and capabilities and, more importantly, begun to define what its limitations might be. The answers to these outstanding questions will be integral to advancing Cas9 as a tool for genome editing with practical and therapeutic applications.

Even so, the use of Cas9 to address biological problems has already begun, spanning from quick, targeted modification of single genes (*12*) for functional validation and disease model generation to large-scale, unbiased genome-wide screens for contributing factors towards a particular phenotype (*13-15*). Pooled CRISPR libraries and multiplexed approaches represent particularly powerful methods that were previously unattainable with either TALENs or ZFNs and provide a more effective knockout-based alternative to genome-wide RNAi screens. Given the overwhelming amount of genomic data now available, the ability of CRISPR to cover both ends of the spectrum from precise target modification to facilitation of robust, sensitive high-throughput screens creates new possibilities in the study of genome structural variations, complex diseases involving multiply interacting pathways, and drug target identification, among others.

Beyond genome modifications

Beginning with the central dogma, much of twentieth century molecular biology has focused on the elucidation of functional genetics. We now know that beyond sequence information, there exists an additional layer of information not encoded within the DNA sequence proper that allows cells carrying identical information to evolve from a monomorphic population to a complex organism. The information encoded in the methylation of DNA or modification of histone proteins account for much of the observable biological variability, and their

aberrant regulation can lead to a wide range of diseases from developmental neuropsychiatric disorders to cancer (*16, 17*).

The identification of the key components involved in creating the different classes of epigenetic markers, together with their chemical inhibitors, have allowed us to gain a basic knowledge of how these marks are established and maintained on a global scale as well as to appreciate how wide-spread, non-specific manipulation of these marks may influence key developmental pathways and cellular functions. Nevertheless, studies regarding the functional consequences or induction of focal epigenetic changes have been largely constrained by the limited specificity, efficiency, and resolution of available technologies, highlighting the need for a set of tools that can more precisely and easily manipulate the epigenome of any organism.

The ease of multiplexed targeting with Cas9 is particularly attractive for targeted epigenome modifications, given that epigenetic regulation is characterized often not by a single mark but a clustering or a series of changes (*18, 19*). Thus, coupling epigenetic effector proteins or their component domains to Cas9 – a natural extension of the growing Cas9 nuclease, dynamic imaging, and transcriptional modulator toolset – might offer a generalizable solution to epigenetic engineering.

A means towards safer gene therapy?

The promise of gene therapy – the idea that defective, disease-causing genes could be replaced by the exogenous transfer of a good copy – was first described over four decades ago (*20*). The development and application of gene therapy was unfortunately marked by two disheartening clinical trial setbacks (*21*) (*22*) that illustrate the major challenges in the field: first, concerns about the immunogenicity-related safety of viral vehicles needed to deliver the genomic material, and second, the specificity of genome editing, where genes other than the intended target might be inadvertently and permanently modified.

The *in vivo* editing of endogenous genes, initiated by ZFNs and TALENs, represents a paradigm shift away from gene replacement therapy approach: rather than transient therapies or integrating additional gene copies at random, this preserves the native dosage of gene expression and natural splicing variants. Currently, several promising studies and trials using designer nucleases are underway (*23*).

Meanwhile, AAV has emerged as a promising non-integrating, non-immunogenic therapeutic vehicle. Because of its small payload limit, however, AAV is largely incompatible with TALENs (*24*). Likewise, the best characterized and most widely used Cas9 to date, SpCas9, remains barely under its packaging limit, leaving little room for promoter choice and none for sgRNA. That the smaller SaCas9 ortholog can mediate mammalian genome editing with efficiencies on par with SpCas9, then, is tremendously exciting. Its size allows for flexibility in the selection of promoters, which in combination with organ-specific AAV serotypes (*25*), can be exploited for further achieving specificity in delivery and expression. In addition to shrinking Cas9s to fit the AAV payload constraint, the development and validation of other novel methods for delivering Cas9 such as safe nanoparticle technologies will be fruitful.

Therapeutics aside, AAV delivery of the CRISPR-Cas system has significant implications for rapid and efficient querying of genetic contributions to both normal biology and disease models beyond cell lines and transgenic animals. The ability to modify somatic tissues with tissue-specificity and temporal control could more accurately model certain processes, such as the post-natal introduction of mutations to simulate the accumulation of genetic insults that may contribute to tumorigenesis. We are thus at a fortuitous juncture of opportunities made available by the increasingly rapid generation of sequencing and bioinformatic data and the development, described in this thesis, of an effective, versatile system for their functional validation: it is my hope then that this ever-expanding toolset will facilitate our ability to address some of the most exciting and challenging questions in biology.

References

1. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)* **337**, 816 (2012).

2. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 86 (2012).

3. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

4. W. Jiang, D. Bikard, D. Cox, F. Zhang, L. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature biotechnology* **31**, 233 (2013).

5. E. Semenova *et al.*, Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10098 (Jun 21, 2011).

6. P. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827 (2013).

7. Y. Fu *et al.*, High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature biotechnology* **31**, 822 (Sep, 2013).

8. N. Crosetto *et al.*, Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature methods* **10**, 361 (Apr, 2013).

9. P. M. Kim, C. Allen, B. M. Wagener, Z. Shen, J. A. Nickoloff, Overexpression of human RAD51 and RAD52 reduces double-strand break-induced homologous recombination in mammalian cells. *Nucleic acids research* **29**, 4352 (Nov 1, 2001).

10. S. Lambert, B. Lopez, Characterization of mammalian RAD51 double strand break repair using non-lethal dominant-negative forms. *The EMBO journal* **19**, 3090 (2000).

11. R. J. Yanez, A. C. Porter, Gene targeting is enhanced in human cells overexpressing hRAD51. *Gene Ther* **6**, 1282 (Jul, 1999).

12. M. N. Lee *et al.*, Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (Mar 7, 2014).

13. O. Shalem *et al.*, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84 (Jan 3, 2014).

14. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80 (Jan 3, 2014).

15. H. Koike-Yusa, Y. Li, E. P. Tan, M. D. Velasco-Herrera, K. Yusa, Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature biotechnology*, (Dec 23, 2013).

16. A. Portela, M. Esteller, Epigenetic modifications and human disease. *Nature biotechnology* **28**, 1057 (Oct, 2010).

17. G. Egger, G. Liang, A. Aparicio, P. A. Jones, Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429**, 457 (May 27, 2004).

18. R. I. Verona, M. R. Mann, M. S. Bartolomei, Genomic imprinting: intricacies of epigenetic regulation in clusters. *Annual review of cell and developmental biology* **19**, 237 (2003).

19. F. Fuks, DNA methylation and histone modifications: teaming up to silence genes. *Current opinion in genetics & development* **15**, 490 (Oct, 2005).

20. T. Friedmann, R. Roblin, Gene therapy for human genetic disease? *Science* **175**, 949 (Mar 3, 1972).

21. S. Hacein-Bey-Abina *et al.*, LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415 (Oct 17, 2003).

22. N. Somia, I. M. Verma, Gene therapy: trials and tribulations. *Nature reviews. Genetics* **1**, 91 (Nov, 2000).

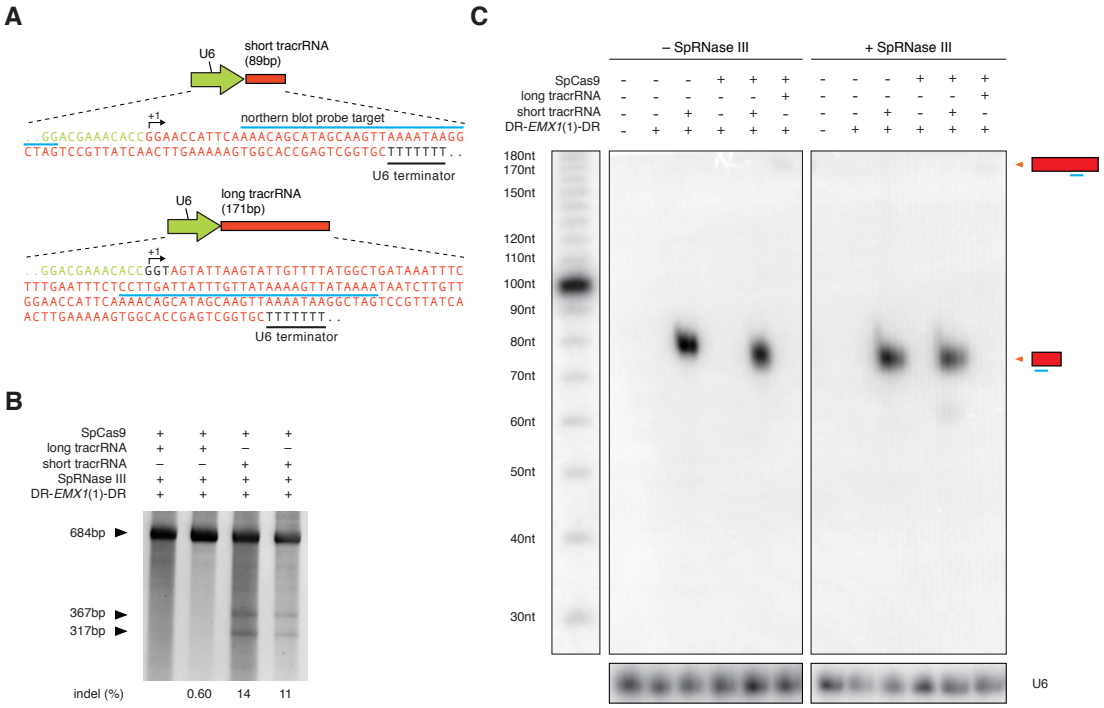
23. T. Gaj, C. Gersbach, C. Barbas, ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology*, (2013).

24. M. Holkers *et al.*, Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic acids research* **41**, e63 (Mar 1, 2013).

25. Z. Wu, A. Asokan, R. J. Samulski, Adeno-associated virus serotypes: vector toolkit for human gene therapy. *Mol Ther* **14**, 316 (Sep, 2006).

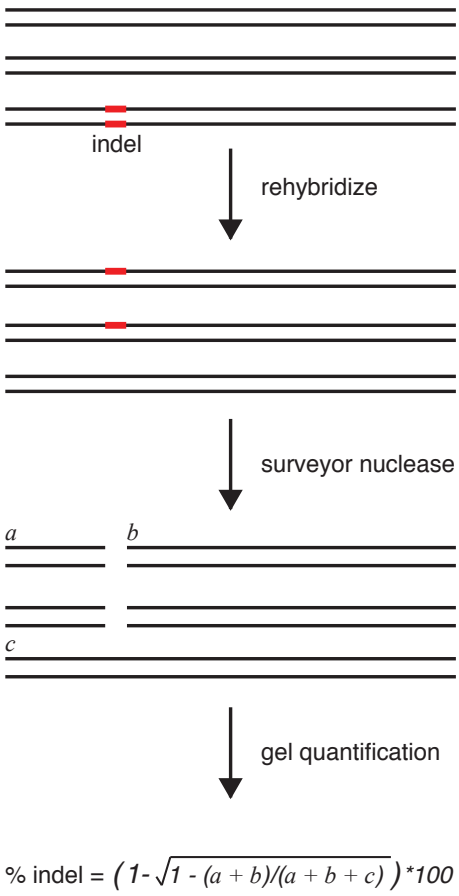
Appendix A Supplementary Figures and Tables

Supplementary Figures



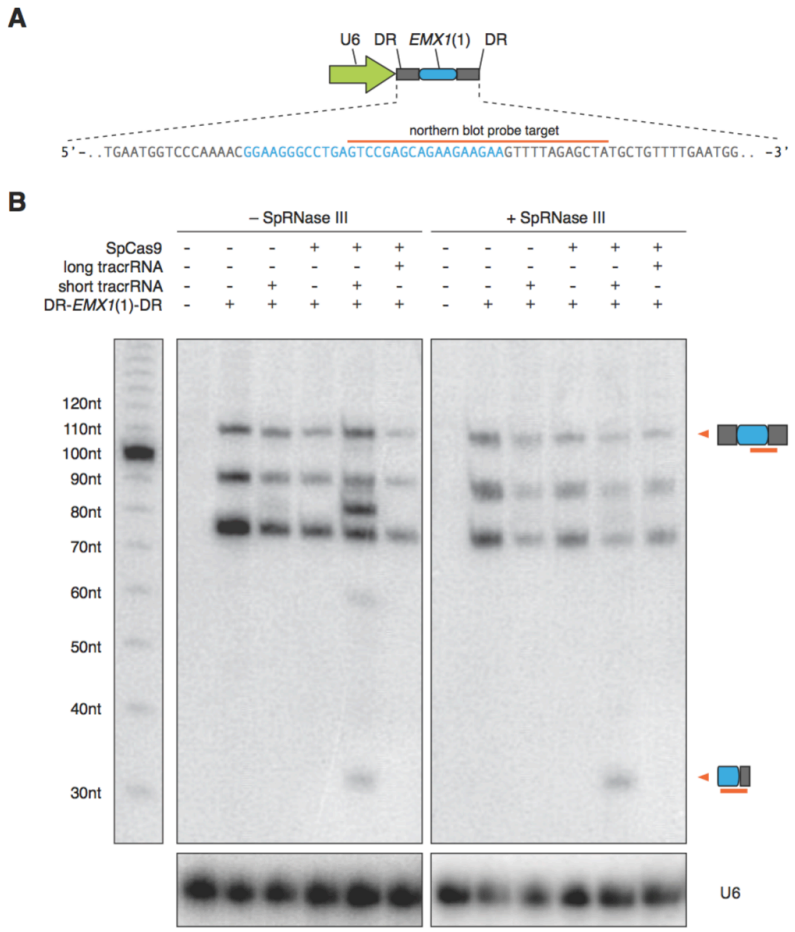
Supplementary Figure 1 Processing of tracrRNA in mammalian cells

(A) Schematic showing the design and sequences of two tracrRNA transcripts tested (short and long). Each transcript is driven by a U6 promoter. Transcription start site is marked as +1 and transcription terminator is as indicated. Blue line indicates the region whose reverse-complement sequence is used to generate northern blot probes for tracrRNA detection. (B) SURVEYOR assay comparing the efficiency of SpCas9-mediated cleavage of the *EMX1* locus. Two biological replicas are shown for each tracrRNA transcript. (C) Northern blot analysis of total RNA extracted from 293FT cells transfected with U6 expression constructs carrying long or short tracrRNA, as well as SpCas9 and DR-*EMX1*(1)-DR. Left and right panels are from 293FT cells transfected without or with SpRNase III respectively. U6 indicate loading control blotted with a probe targeting human U6 snRNA. Transfection of the short tracrRNA expression construct led to abundant levels of the processed form of tracrRNA (~75bp) (*11*). Very low amounts of long tracrRNA are detected on the northern blot. As a result of these experiments, we chose to use short tracrRNA for application in mammalian cells.



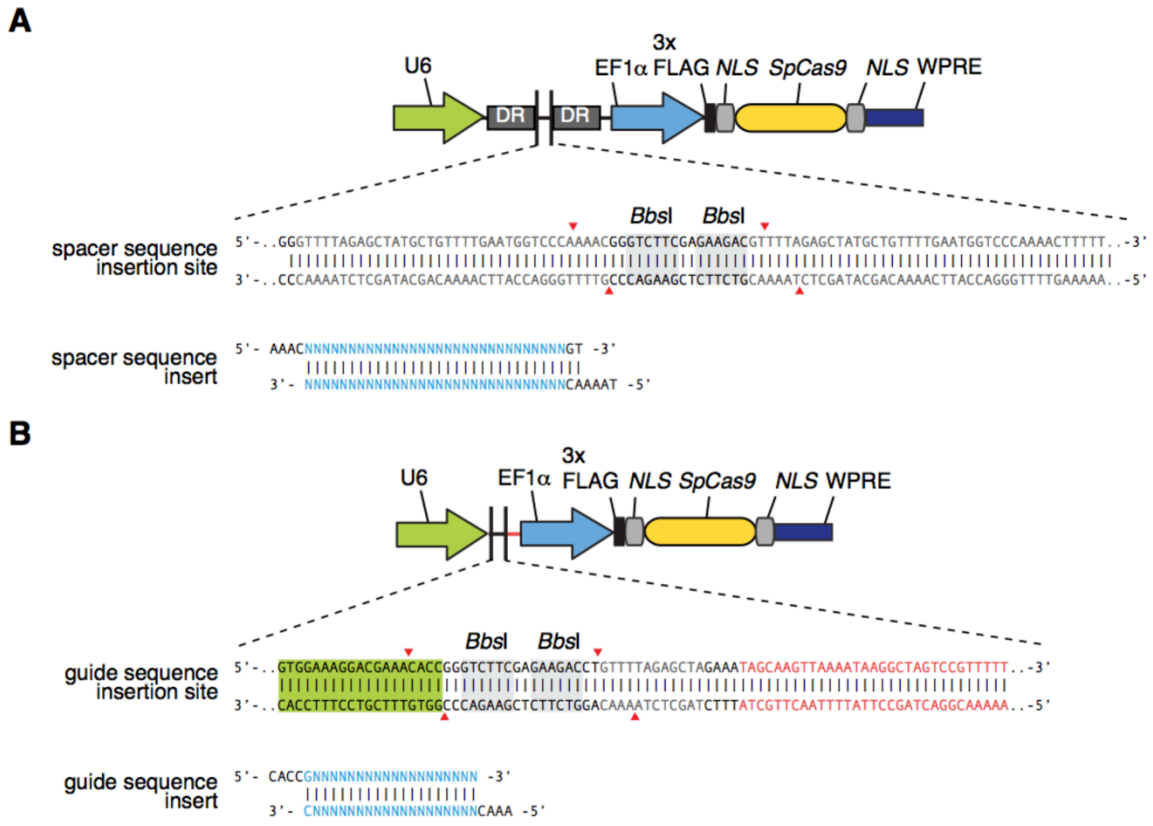
Supplementary Figure 2 Schematic of SURVEYOR assay

Schematic of the SURVEYOR assay used to determine Cas9-mediated cleavage efficiency. First, genomic PCR (gPCR) is used to amplify the Cas9 target region from a heterogeneous population of modified and unmodified cells, and the gPCR products are reannealed slowly to generate heteroduplexes. The reannealed heteroduplexes are cleaved by SURVEYOR nuclease, whereas homoduplexes are left intact. Cas9-mediated cleavage efficiency (% indel) is calculated based on the fraction of cleaved DNA.



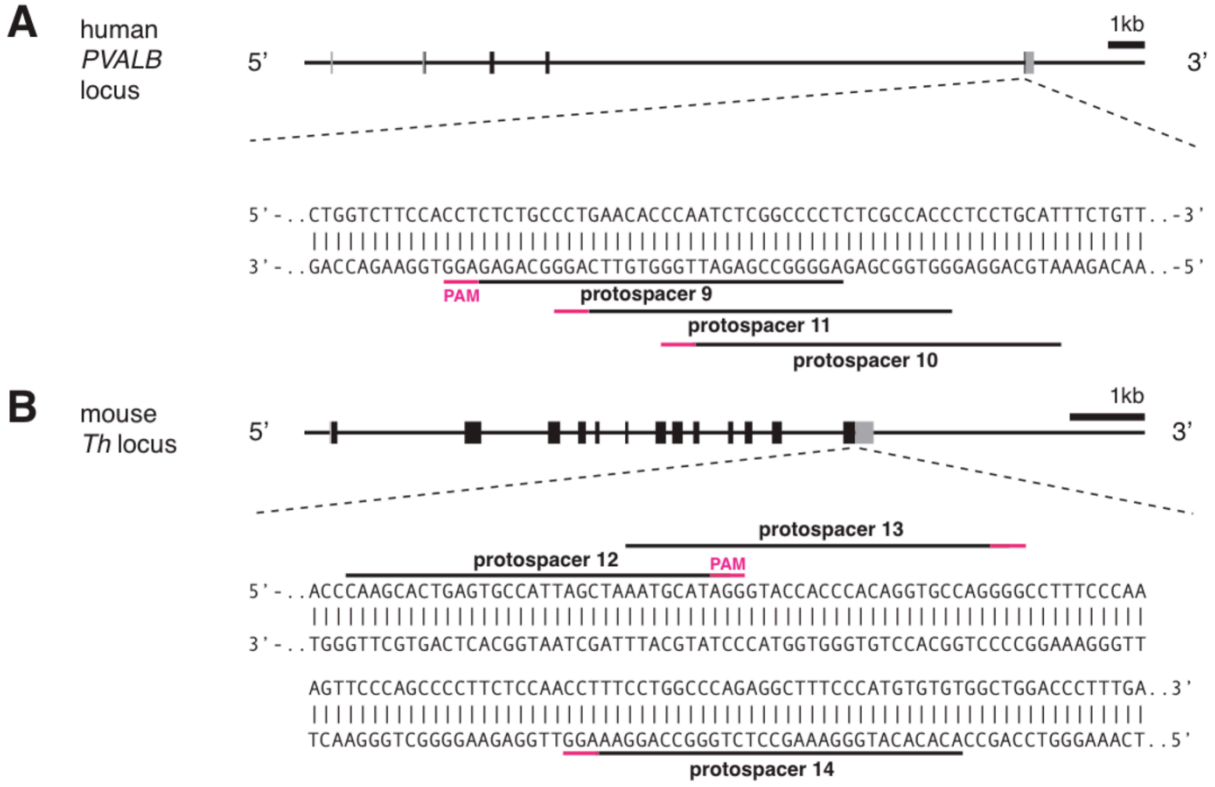
Supplementary Figure 3 Processing of crRNA in mammalian cells

(A) Schematic showing the expression vector for a single spacer flanked by two direct repeats (DR-*EMX1(1)-DR*). The 30bp spacer targeting the human *EMX1* locus protospacer 1 (Supplementary Table 1) is shown in blue and direct repeats are in shown in gray. Orange line indicates the region whose reverse- complement sequence is used to generate northern blot probes for *EMX1*(1) crRNA detection. (B) Northern blot analysis of total RNA extracted from 293FT cells transfected with U6 expression constructs carrying DR-*EMX1*(1)-DR. Left and right panels are from 293FT cells transfected without or with SpRNase III respectively. DR-*EMX1*(1)-DR was processed into mature crRNAs only in the presence of SpCas9 and short tracrRNA, and was not dependent on the presence of SpRNase III. The mature crRNA detected from transfected 293FT total RNA is ~33bp and is shorter than the 39-42bp mature crRNA from *S. pyogenes* (*19*), suggesting that the processed mature crRNA in human 293FT cells is likely different from the bacterial mature crRNA in *S. pyogenes*.



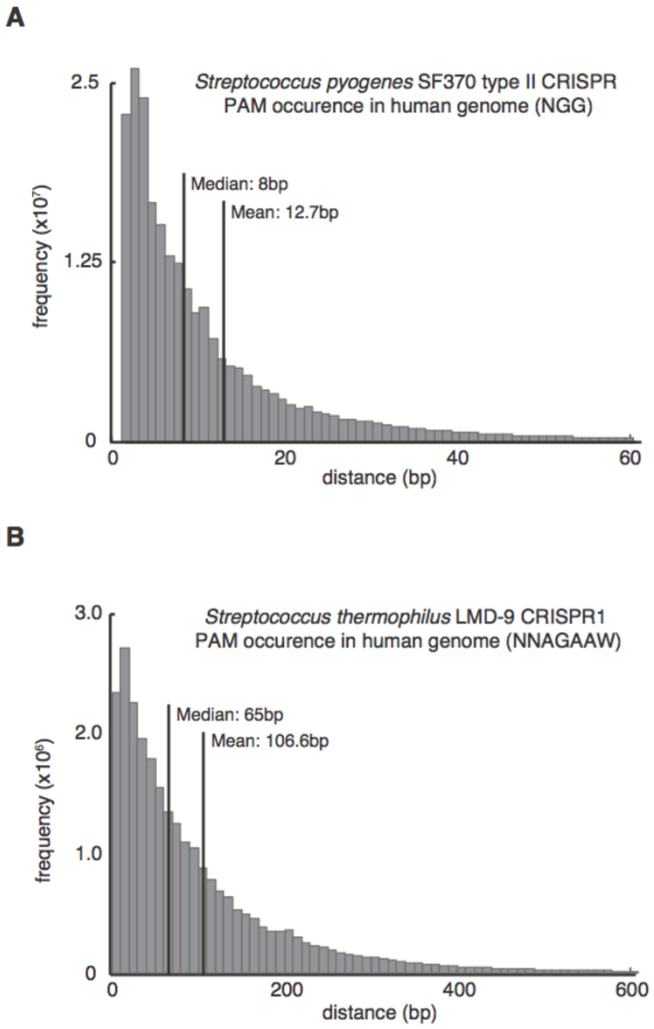
Supplementary Figure 4 Bi-cistronic expression vectors for crRNA or chimeric RNA and SpCas9

(A) Schematic showing the design of an expression vector for the pre-crRNA array. Spacers can be inserted between two *Bbs*I sites using annealed oligonucleotides. Sequence design for the oligonucleotides are shown below with the appropriate ligation adapters indicated. (B) Schematic of the expression vector for chimeric crRNA. The guide sequence can be inserted between two *Bbs*I sites using annealed oligonucleotides. The vector already contains the partial direct repeat (gray) and partial tracrRNA (red) sequences. WPRE, Woodchuck hepatitis virus post- transcriptional regulatory element.



Supplementary Figure 5 Selection of loci in human and mouse cells

Schematic of the human *PVALB* (A) and mouse *Th* (B) loci and the location of the three protospacers within the last exon of the *PVALB* and *Th* genes, respectively. The 30-bp protospacers are indicated by black lines and the adjacent PAM sequences are indicated by the magenta bar. Protospacers on the sense and anti-sense strands are indicated above and below the DNA sequences respectively.

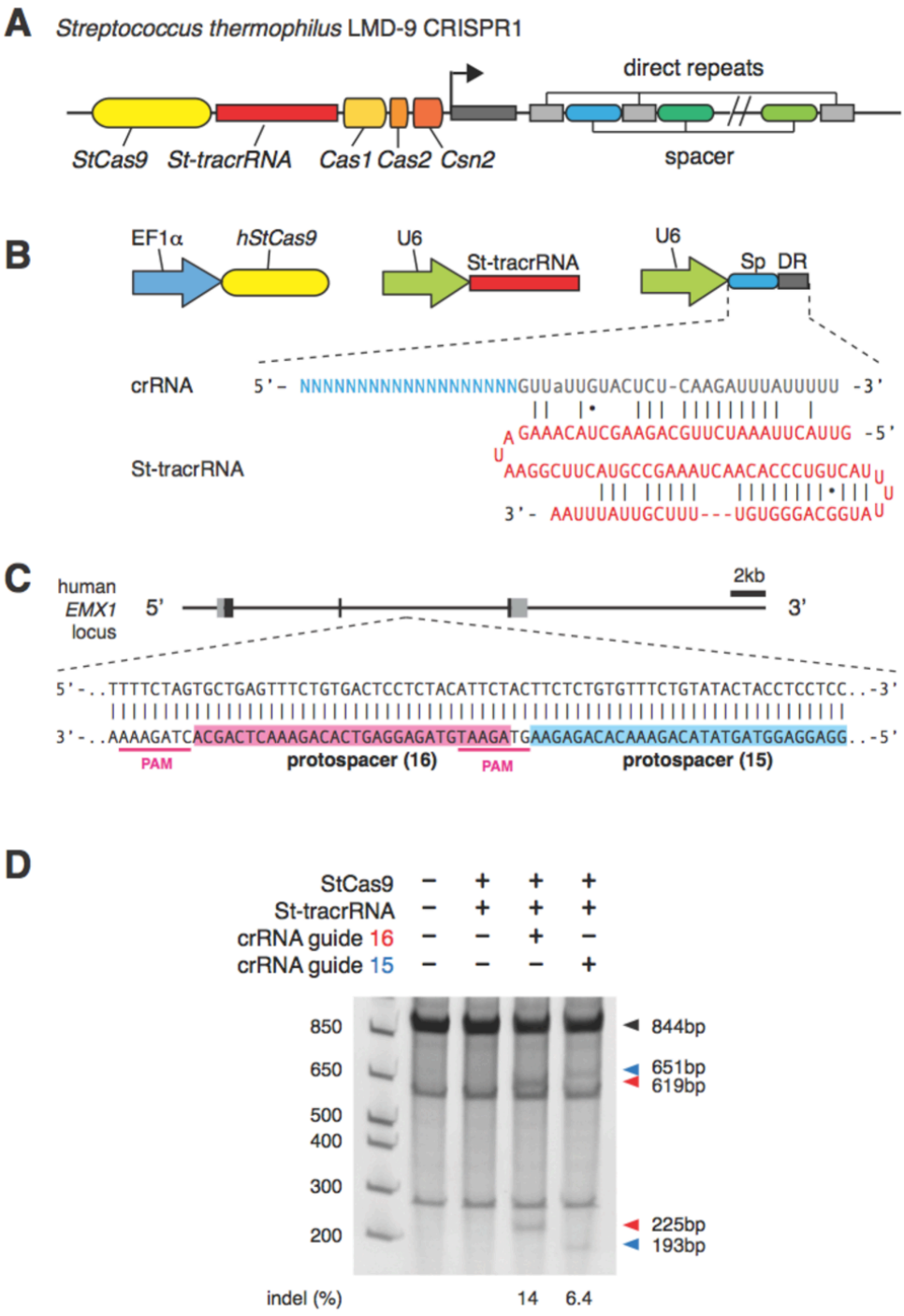


C

	NGG		NNAGAAW	
Chr	median	mean	median	mean
1	7	12.8	67	115.8
2	8	12.7	64	100.8
3	8	13.0	63	98.5
4	9	14.0	61	94.5
5	8	13.1	63	97.9
6	8	13.1	63	98.5
7	8	12.4	64	102.9
8	8	12.8	64	100.9
9	7	13.9	65	120.5
10	7	12.1	66	107.0
11	7	12.0	65	105.8
12	8	12.4	65	103.5
13	8	13.6	62	94.6
14	8	12.0	65	101.5
15	7	11.5	68	107.7
16	7	11.7	74	136.8
17	6	10.3	76	127.9
18	8	13.4	63	101.8
19	6	9.4	82	145.4
20	7	11.1	72	121.8
21	7	13.4	64	111.4
22	6	9.2	85	140.3
X	8	13.2	63	99.0
Y	8	29.2	62	223.7

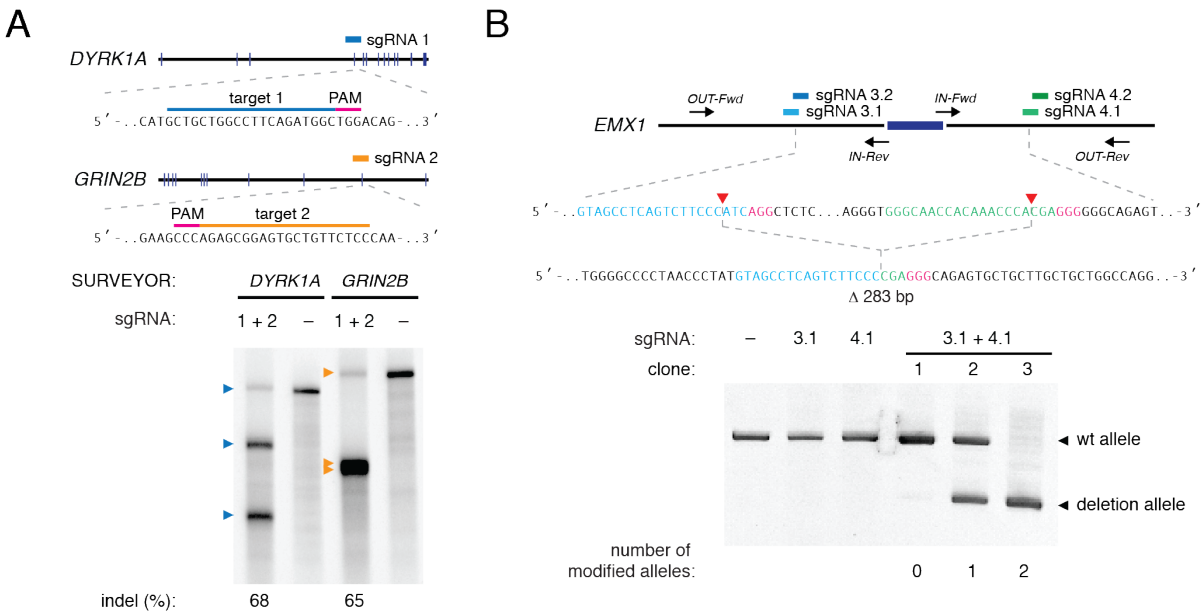
Supplementary Figure 6 Distribution frequency of PAM in the human genome

Histograms of distances between adjacent *Streptococcus pyogenes* SF370 type II CRISPR PAM (NGG) (A) and *Streptococcus thermophiles* LMD-9 CRISPR1 PAM (NNAGAAW) (B) in the human genome. (C) Distances for each PAM by chromosome. Chr, chromosome. Putative targets were identified using both the plus and minus strands of human chromosomal sequences. Given that there may be chromatin, DNA methylation-, RNA structure, and other factors that may limit the cleavage activity at some protospacer targets, it is important to note that the actual targeting ability might be less than the result of this computational analysis.



Supplementary Figure 7 Mammalian gene targeting using S. thermophilus CRISPR1

(A) Schematic of CRISPR locus 2 from *Streptococcus thermophilus* LMD-9. (B) Design of the expression system for the *S. thermophilus* CRISPR system. Human codon- optimized *StCas9* is expressed using a constitutive EF1a promoter. Mature versions of tracrRNA and crRNA are expressed using the U6 promoter to ensure precise transcription initiation. Sequences for the mature crRNA and tracrRNA are shown. A single based indicated by the lower case "a" in the crRNA sequence was used to remove the polyU sequence, which serves as a RNA Pol III transcriptional terminator. Sp, spacer. (C) Schematic showing protospacer and corresponding PAM sequences targets in the human *EMX1* locus. Two protospacer sequences are highlighted and their corresponding PAM sequences satisfying the NNAGAAW motif are indicated by magenta lines. Both protospacers are targeting the anti-sense strand. (D) SURVEYOR assay showing StCas9-mediated cleavage in the target locus. RNA guide spacers 1 and 2 induced 14% and 6.4% respectively. Statistical analysis of cleavage activity across biological replica at these two protospacer sites can be found in Supplementary Table 1.

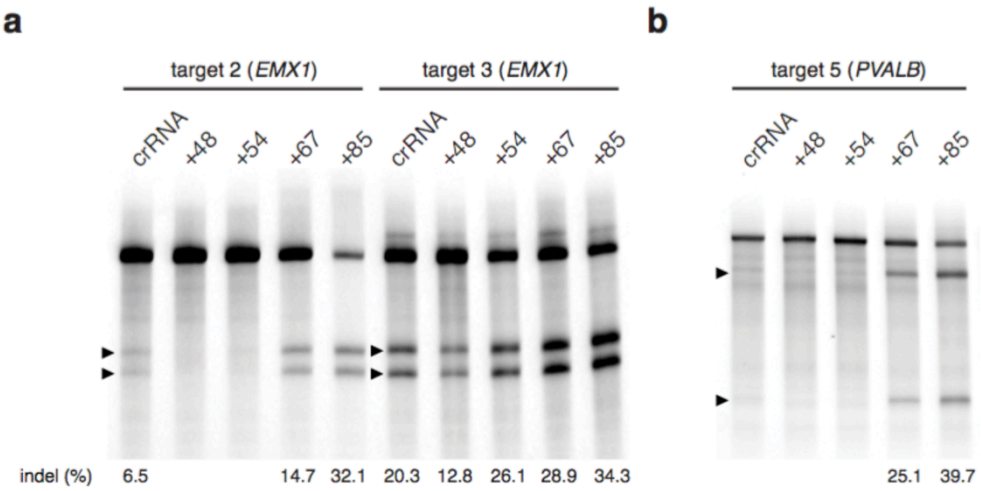


C

	Approx. deletion	Number of clones				
sgRNAs	*size (bp)*	*+/+*	*Δ/+*	*Δ/Δ*	*inversion*	*Total*
3.1 + 4.1	282	12	10	1	0	23
3.2 + 4.1	237	20	16	2	0	38
3.1 + 4.2	425	4	14	5	0	23
3.2 + 4.2	259	6	11	4	0	21

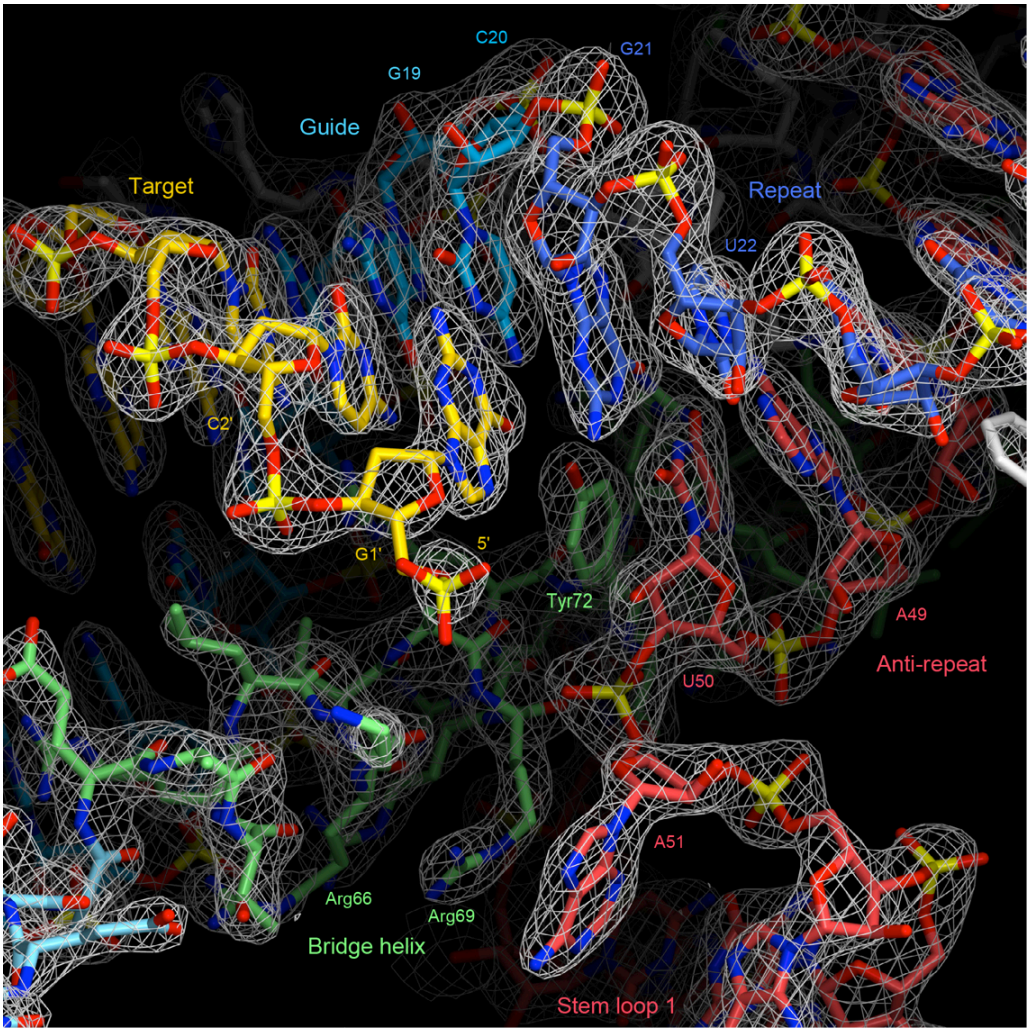
Supplementary Figure 8 Multiplexed gene targeting and microdeletion with optimized sgRNAs

(A) Multiplexed targeting of endogenous human genes using co-transfection of Cas9 and PCR amplicons containing U6-driven sgRNAs. (B) Microdeletion of an exon in the human *EMX1* gene mediated by a pair of sgRNAs. Gel shows genotypes of representative transfected cells after clonal isolation and expansion, with no modification, mono-allelic deletion, and bi-allelic deletion. (C) Table accounting for the frequency of modifications using pairs of sgRNAs inducing microdeletions.



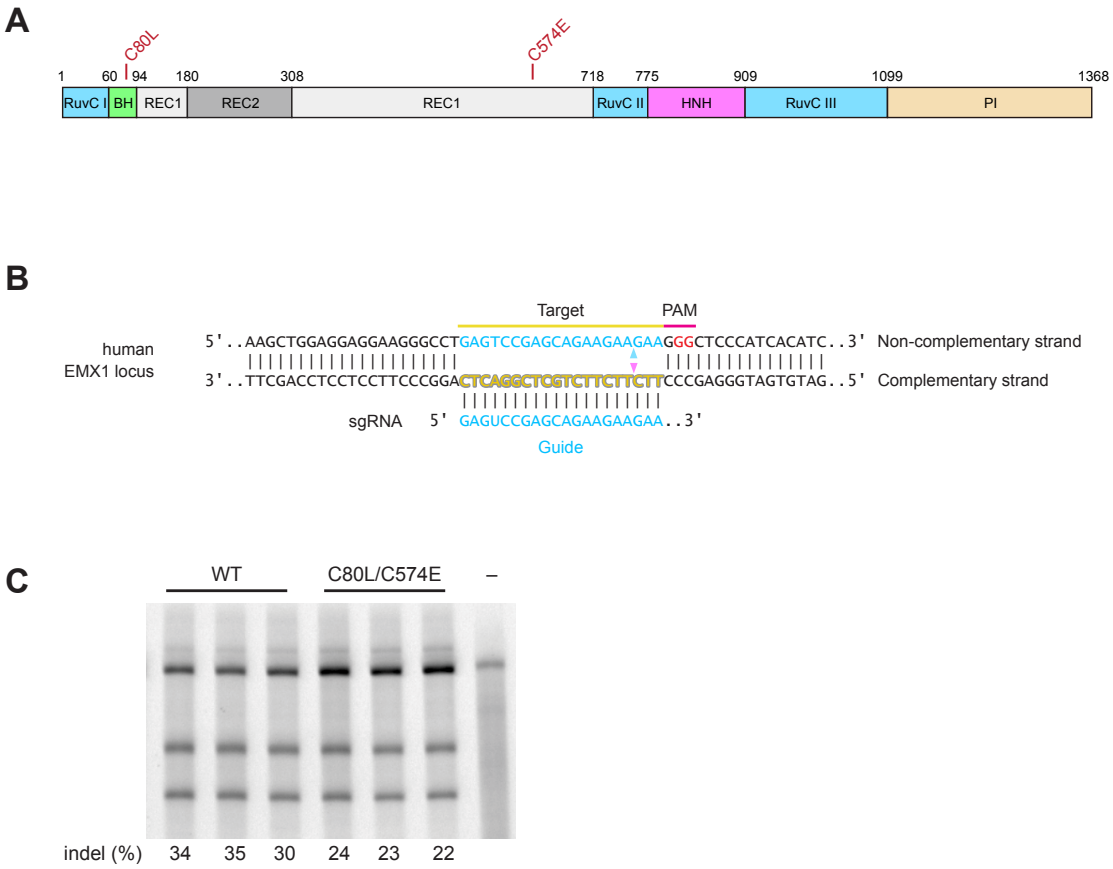
Supplementary Figure 9 Optimization of sgRNA architecture

DNA expression vectors carrying SpCas9 and crRNA-tracrRNA pair or single guide RNA (sgRNA) are co-transfected into 293FT cells. Cleavage efficiency (% indel) is assessed using the SURVEYOR nuclease assay as described. Modification efficiencies at (A) 2 *EMX1* loci and (B) 1 *PVALB* locus are shown. Arrows indicate the expected SURVEYOR fragments.



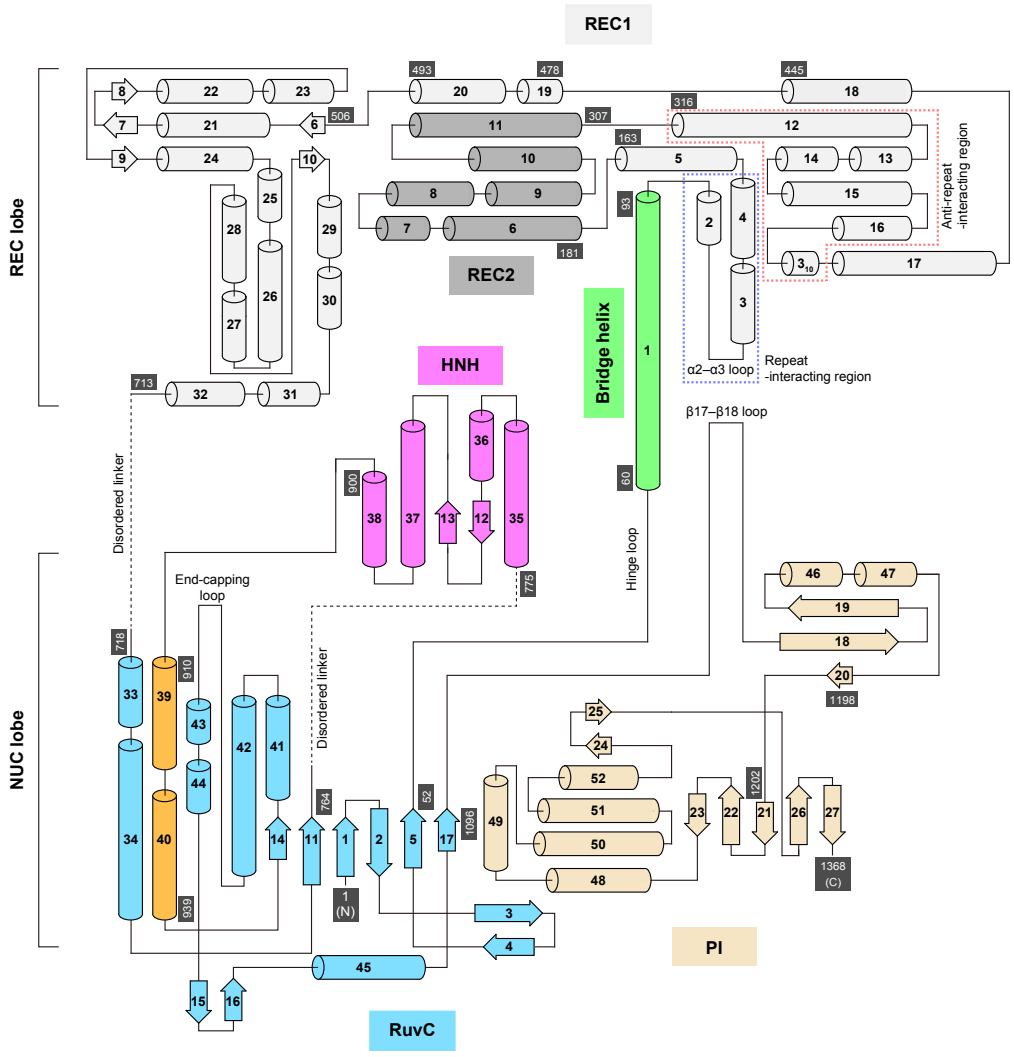
Supplementary Figure 10 Electron density map

The $2mF_O - DF_C$ electron density map around the three-way junction is shown as a gray mesh (contoured at 2.5σ).

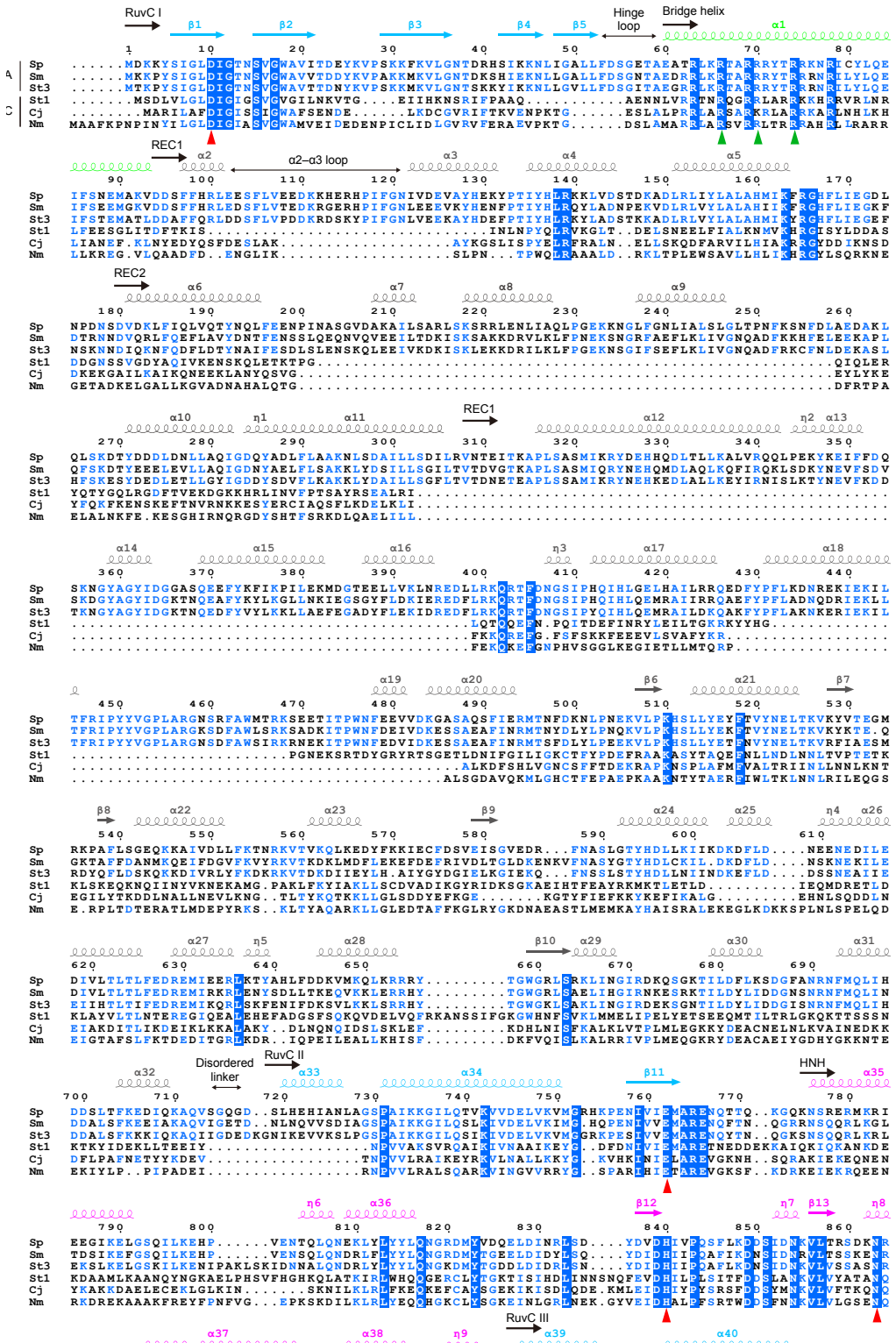


Supplementary Figure 11 The di-cysteine mutant (C80L/C574E) is functional in HEK 293FT cells

(A) Schematic illustration of the positions of the cysteine mutations (C80L and C574E) in Cas9. (B) Sequence of the target site used to test the function of the C80L/C574E mutant of Cas9. The cleavage sites by the HNH and RuvC domains are indicated by pink and cyan triangles, respectively. (C) SURVEYOR nuclease assay, showing indels generated by either the wild-type or C80L/C574E mutant (n = 3).



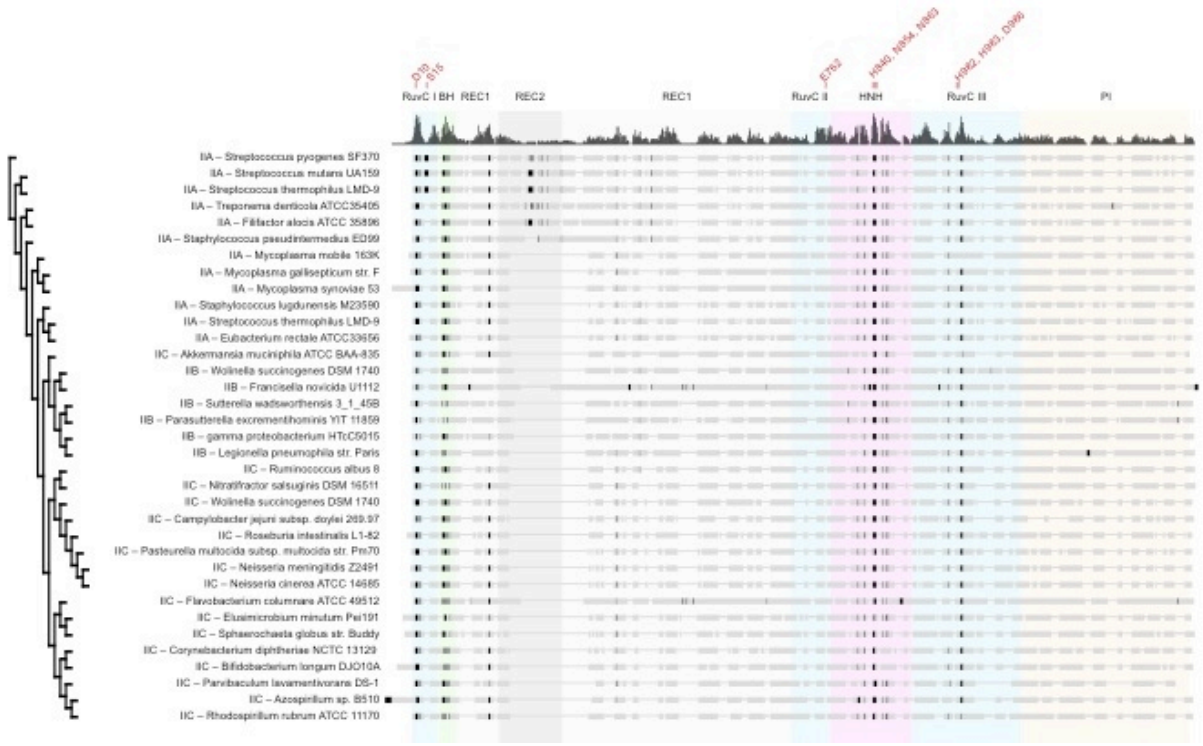
Supplementary Figure 12 Schematic drawing of the secondary structural element of Cas9


RuvC I
β1
β2
β3
β4
β5
Hinge loop
Bridge helix
α1
A
C
Sp
Sm
St3
St1
Cj
Nm
REC1
α2
α2–α3 loop
α3
α4
α5
REC2
α6
α7
α8
α9
α10
η1
α11
REC1
α12
η2
α13
α14
α15
α16
η3
α17
α18
α19
α20
β6
α21
β7
β8
α22
α23
β9
α24
α25
η4
α26
α27
η5
α28
β10
α29
α30
α31
Disordered linker
RuvC II
α32
α33
α34
β11
HNH
α35
η6
α36
β12
η7
β13
η8
α37
α38
η9
RuvC III
α39
α40

(continued)

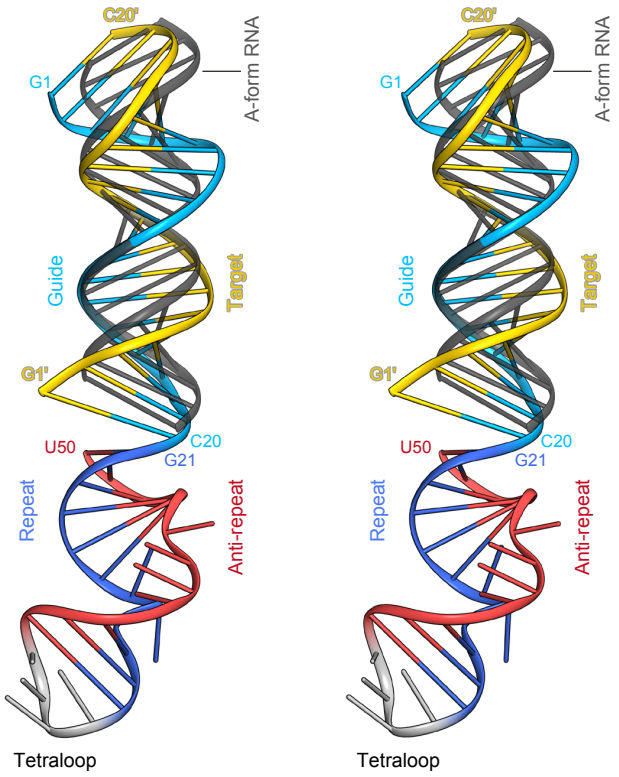
Supplementary Figure 13 Sequence alignment of Cas9 orthologs in families II-A and II-C

The catalytic residues are indicated by red triangles. Critical arginine residues on the Bridge helix are indicated by green triangles. The secondary structure of *S. pyogenes* Cas9 is shown above the sequences. The figure was prepared using ClustalW (*12*) and ESPript (*14*). Sp, *S. pyogenes*; Sm, *Streptococcus mutans*; St3, *Streptococcus thermophilus* CRISPR-3; St1, *Streptococcus thermophilus* CRISPR-1; Cj, *Campylobacter jejuni*; Nm, *Neisseria meningitidis*.



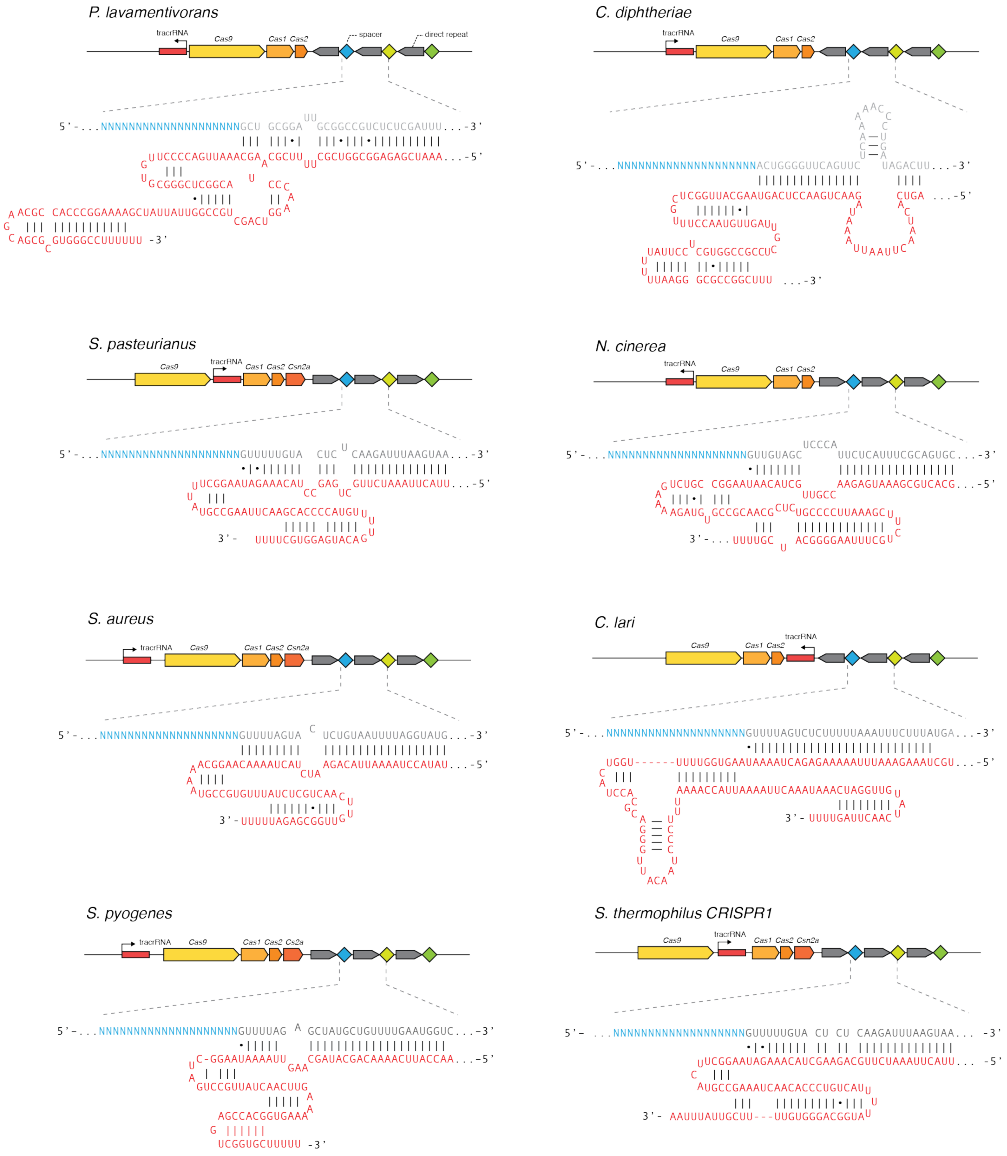
Supplementary Figure 14 Sequence alignment of Cas9 orthologs in families II-A, II-B, and II-C

Thirty-five Cas9 orthologs from families IIA, IIB and IIC were aligned (BLOSUM62) and clustered (Jukes-Cantor model Neighbor-Joining method, with *S. pyogenes* Cas9 as the outgroup). Bars on the top show amino acid conservation. In each line, the black bars show residues with at least 75% consensus, and the gray bars show non-conserved residues.



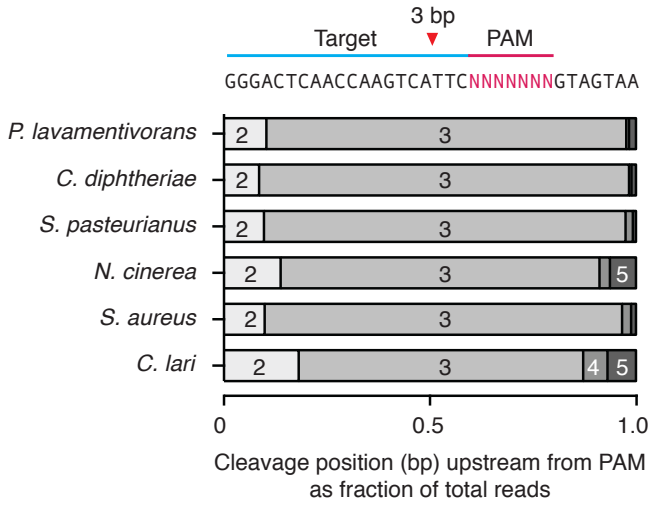
Supplementary Figure 15 Comparison of the guide:target heteroduplex with a canonical A-form RNA duplex

The guide:target heteroduplex was superimposed on an A-form RNA duplex, based on their phosphorus atoms. The A-form RNA duplex is colored dark gray. Nucleotides 51–97 of the sgRNA were omitted, for clarity.



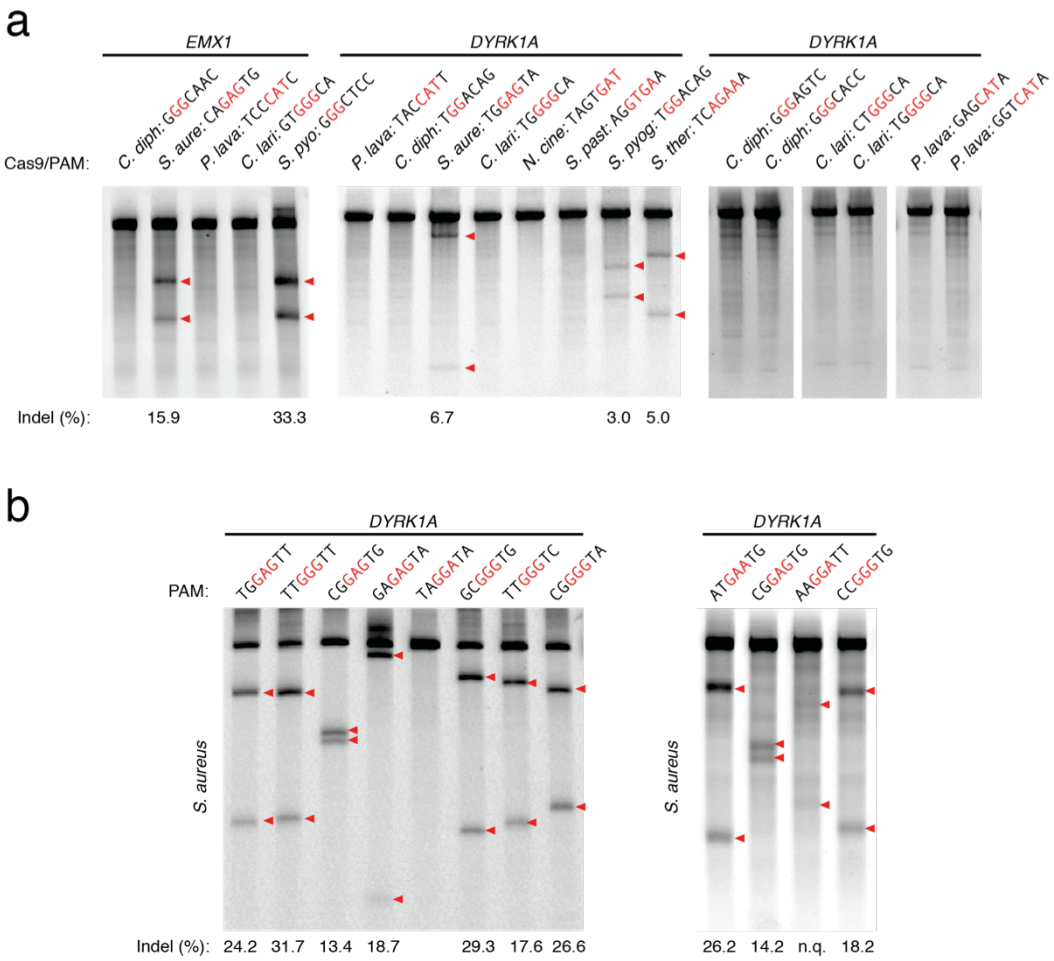
Supplementary Figure 16 Schematic of Type II CRISPR loci for orthologs

Spacer or "guide" sequences are shown in blue, followed by direct repeat (gray). Predicted tracrRNAs are shown in red, and folded based on the Constraint Generation RNA folding model[6].



Supplementary Figure 17 Cleavage position of Cas9 orthologs

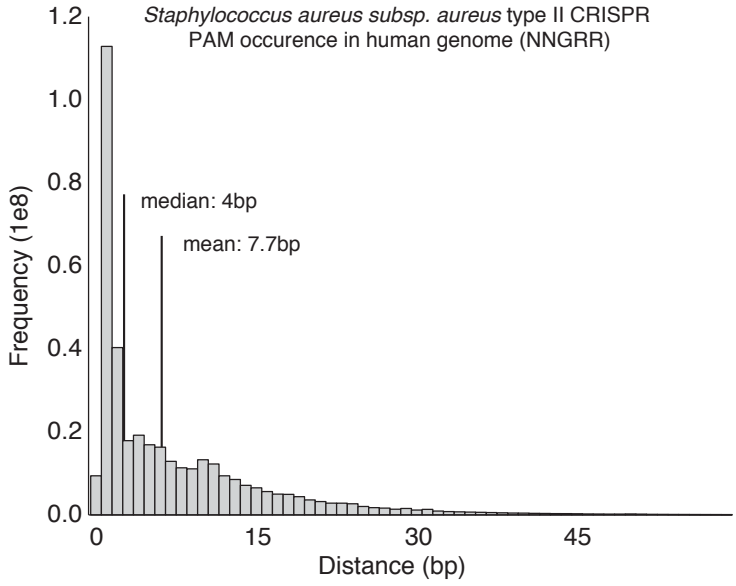
Stacked bar graph indicates the fraction of targets cleaved at 2, 3, 4, or 5-bp upstream of PAM for each Cas9 ortholog; all Cas9s cleave most frequently at 3-bp upstream of PAM (red triangle).



Supplementary Figure 18 Mammalian endogenous gene targeting by Cas9 orthologs

(A) SURVEYOR assays showing indel formation at human endogenous loci from co-transfection of Cas9 orthologs and sgRNA in HEK 293FT cells. (B) SaCas9 cleaves multiple targets with high efficiency. PAM sequences for individual targets are shown above each lane, with consensus sequences for each Cas9 highlighted in red. Red triangles indicate cleaved fragments.

a

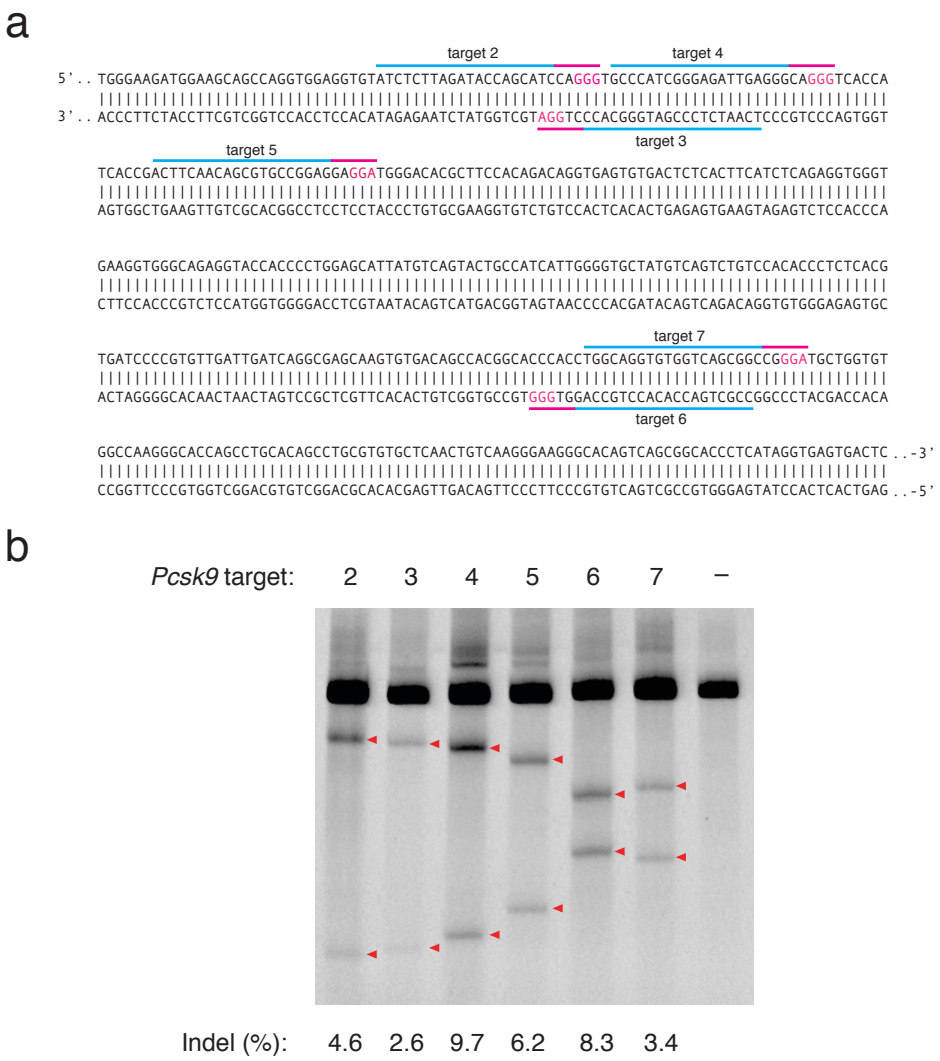


b

	NNGRR	
Chr	median	mean
1	4	7.5
2	4	7.9
3	4	7.9
4	4	8.3
5	4	8.0
6	4	7.9
7	4	7.7
8	4	7.8
9	4	7.7
10	4	7.5
11	4	7.5
12	4	7.7
13	4	8.2
14	4	7.7
15	4	7.5
16	4	7.0
17	4	6.8
18	4	7.8
19	3	6.5
20	4	7.1
21	4	8.0
22	3	6.7
X	4	8.0
Y	4	8.1

Supplementary Figure 19 PAM distribution frequency for SaCas9

(A) Histograms of distances between adjacent *Staphylococcus aureus subsp. aureus* Type II CRISPR PAM (NNGRR) in the human genome (GRCh38). (B) Distances for each PAM by chromosome.



Supplementary Figure 20 SaCas9 target loci for Pcsk9 gene

(A) Location of SaCas9 targets and PAMs within the mouse *Pcsk9* gene locus. (B) Indels produced at target sites from transfection of mouse liver hepatoma (Hepa1-6) cell line. Red arrows indicate cleavage sites.

Supplementary Tables

Supplementary Table 1 Human and mouse loci targeted in the study

Cas9	target species	gene	protospacer ID	protospacer sequence (5' to 3')	PAM	strand	cell line tested	% indel (pre-crRNA + tracrRNA)	% indel (chimeric RNA)
S. pyogenes SF370 type II CRISPR	*Homo sapiens*	*EMX1*	1	GGAAGGGCCTGAGTCCGAGCAGAAGAAGAA	GGG	+	293FT	20 ± 1.8	6.7 ± 0.62
		EMX1	2	CATTGGAGGTGACATCGATGTCCTCCCCAT	TGG	–	293FT	2.1 ± 0.31	N.D.
		EMX1	3	GGACATCGATGTCACCTCCAATGACTAGGG	TGG	+	293FT	14 ± 1.1	N.D.
		EMX1	4	CATCGATGTCCTCCCCATTGGCCTGCTTCG	TGG	–	293FT	11 ± 1.7	N.D.
		EMX1	5	TTCGTGGCAATGCGCCACCGGTTGATGTGA	TGG	–	293FT	4.3 ± 0.46	2.1 ± 0.51
		EMX1	6	TCGTGGCAATGCGCCACCGGTTGATGTGAT	GGG	–	293FT	4.0 ± 0.66	0.41 ± 0.25
		EMX1	7	TCCAGCTTCTGCCGTTTGTACTTTGTCCTC	CGG	–	293FT	1.5 ± 0.12	N.D.
		EMX1	8	GGAGGGAGGGGCACAGATGAGAAACTCAGG	AGG	–	293FT	7.8 ± 0.83	2.3 ± 1.2
	Homo sapiens	*PVALB*	9	AGGGGCCGAGATTGGGTGTTCAGGGCAGAG	AGG	+	293FT	21 ± 2.6	6.5 ± 0.32
		PVALB	10	ATGCAGGAGGGTGGCGAGAGGGGCCGAGAT	TGG	+	293FT	N.D.	N.D.
		PVALB	11	GGTGGCGAGAGGGGCCGAGATTGGGTGTTC	AGG	+	293FT	N.D.	N.D.
	Mus musculus	*Th*	12	CAAGCACTGAGTGCCATTAGCTAAATGCAT	AGG	–	Neuro2A	27 ± 4.3	4.1 ± 2.2
		Th	13	AATGCATAGGGTACCACCCACAGGTGCCAG	GGG	–	Neuro2A	4.8 ± 1.2	N.D.
		Th	14	ACACACATGGGAAAGCCTCTGGGCCAGGAA	AGG	+	Neuro2A	11.3 ± 1.3	N.D.
S. thermophilus LMD-9 CRISPR1	*Homo sapiens*	*EMX1*	15	GGAGGAGGTAGTATACAGAAACACAGAGAA	GTAGAAT	–	293FT	14 ± 0.88	N.T.
		EMX1	16	AGAATGTAGAGGAGTCACAGAAACTCAGCA	CTAGAAA	–	293FT	7.8 ± 0.77	N.T.

Protospacer targets designed based on *Streptococcus pyogenes* type II CRISPR and *Streptococcus thermophilus* CRISPR1 loci with their requisite PAMs against three different genes in human and mouse genomes. Cells were transfected with Cas9 and either pre- crRNA/tracrRNA or chimeric RNA. Cells were analyzed 72 hours after transfection. Percent indels are calculated based on SURVEYOR assay results from indicated cell lines, $N = 3$ for all protospacer targets, errors are S.E.M. N.D., not detectable using the SURVEYOR assay; N.T., not tested in this study.

Supplementary Table 2 List of paired sgRNAs used in the study

Gene	Overhang Length (bp)	Overhang Type	Offset Length (bp)	Cas9n with both sgRNAs indel (%)	Top strand-nicking sgRNA target site		Top strand-sgRNA with WT Cas9 indel (%)	Bottom strand-nicking sgRNA target site		Bottom strand-sgRNA with WT Cas9 indel (%)
					Guide sequence (5′ to 3′)	PAM		Guide sequence (5′ to 3′)	PAM	
EMX1	147	3′	-181	N.D.	TGCGCCACCGGTTGATGTGA	TGG	13.15	AGGCCCCAGTGGCTGCTCTG	GGG	27.29
EMX1	100	3′	-134	N.D.	ACTCTGCCCTCGTGGGTTTG	TGG	24.7	GAGTCCGAGCAGAAGAAGAA	GGG	21.9
EMX1	47	3′	-81	N.D.	ACTCTGCCCTCGTGGGTTTG	TGG	24.7	CACGAAGCAGGCCAATGGGG	AGG	13.57
EMX1	24	3′	-58	N.D.	GGAGCCCTTCTTCTTCTGCT	CGG	33.17	CAAACGGCAGAAGCTGGAGG	AGG	26.15
EMX1	16	3′	-50	N.D.	GGGGCACAGATGAGAAACTC	AGG	26.56	AGGCCCCAGTGGCTGCTCTG	GGG	27.29
EMX1	8	3′	-42	N.D.	GCCGTTTGTACTTTGTCCTC	CGG	30.49	TGAAGGTGTGGTTCCAGAAC	CGG	36.02
EMX1	26	5′	-8	13.7 ± 1.27	GCCGTTTGTACTTTGTCCTC	CGG	9.82	CAAACGGCAGAAGCTGGAGG	AGG	26.15
EMX1	30	5′	-4	19.72 ± 0.32	GCCGTTTGTACTTTGTCCTC	CGG	30.49	CGGCAGAAGCTGGAGGAGGA	AGG	22.06
EMX1	31	5′	-3	21.35 ± 2.23	TCACCTGGGCCAGGGAGGGA	GGG	10.75	TGAAGGTGTGGTTCCAGAAC	CGG	36.02
EMX1	34	5′	0	26.89 ± 1.54	TGCGCCACCGGTTGATGTGA	TGG	13.15	TTGCCACGAAGCAGGCCAAT	GGG	13.77
EMX1	38	5′	4	36.31 ± 2.97	TGCGCCACCGGTTGATGTGA	TGG	14.49	CACGAAGCAGGCCAATGGGG	AGG	13.57
EMX1	51	5′	17	31.12 ± 0.25	GGGGCACAGATGAGAAACTC	AGG	26.56	TGAAGGTGTGGTTCCAGAAC	CGG	36.02
EMX1	54	5′	20	32.41 ± 3.68	GGGGCACAGATGAGAAACTC	AGG	26.56	AGGTGTGGTTCCAGAACCGG	AGG	35.53
EMX1	65	5′	31	13.45 ± 1.99	TCACCTGGGCCAGGGAGGGA	GGG	10.75	CAAACGGCAGAAGCTGGAGG	AGG	26.15
EMX1	69	5′	35	12.39 ± 1.29	TCACCTGGGCCAGGGAGGGA	GGG	10.75	CGGCAGAAGCTGGAGGAGGA	AGG	22.06
EMX1	76	5′	42	21.71 ± 1.66	GCCGTTTGTACTTTGTCCTC	CGG	30.49	AGGGCTCCCATCACATCAAC	CGG	41.27
EMX1	85	5′	51	21.89 ± 1.88	GGGGCACAGATGAGAAACTC	AGG	26.56	CAAACGGCAGAAGCTGGAGG	AGG	26.15
EMX1	95	5′	61	5.88 ± 1.81	TCACCTGGGCCAGGGAGGGA	GGG	10.75	TGAGTCCGAGCAGAAGAAGA	AGG	29.06
EMX1	135	5′	101	15.78 ± 2.19	GGGGCACAGATGAGAAACTC	AGG	26.56	AGGGCTCCCATCACATCAAC	CGG	41.27
EMX1	145	5′	111	N.D.	TCACCTGGGCCAGGGAGGGA	GGG	10.75	TTGCCACGAAGCAGGCCAAT	GGG	13.77

(Supp Table 2 Continued)

EMX1	181	5'	147	N.D.	TCACCTGGGCCA GGGAGGGA	GGG	10.75	TCACCTCCAAT GACTAGGGT	GGG	25.14
EMX1	201	5'	167	N.D.	GGGGCACAGATG AGAAACTC	AGG	26.56	TCACCTCCAAT GACTAGGGT	GGG	25.14
EMX1	222	5'	188	N.D.	TCACCTGGGCCA GGGAGGGA	GGG	10.75	GGCAGAGTGCT GCTTGCTGC	TGG	10.75
EMX1	242	5'	208	N.D.	GGGGCACAGATG AGAAACTC	AGG	26.56	GGCAGAGTGCT GCTTGCTGC	TGG	17.22
DYRK1A	163	3'	-197	N.D.	ATCTGGTCAGAA TATGATAA	AGG	10.65 ± 2.05	AACCTCACTTA TCTTCTTGT	AGG	19.02 ± 4.32
DYRK1A	104	3'	-138	N.D.	GTCACTGTACTG ATGTGAAT	TGG	16.71 ± 2.47	AACCTCACTTA TCTTCTTGT	AGG	17.04 ± 1.30
DYRK1A	65	3'	-99	N.D.	CATCTGAAGGCC AGCAGCAT	TGG	8.82 ± 1.01	CTCACTTATCT TCTTGTAGG	AGG	18.79 ± 2.71
DYRK1A	24	3'	-58	N.D.	GTCACTGTACTG ATGTGAAT	TGG	17.83 ± 0.43	CCATGCTGCTG GCCTTCAGA	TGG	17.15 ± 3.29
DYRK1A	3	3'	-37	N.D.	TGATAAGGCAGA AACCTGTT	TGG	4.95 ± 0.66	GCCAAACATAA GTGACCAAC	AGG	16.38 ± 3.39
DYRK1A	18	5'	-16	N.D.	GAAGATAAGTGA GGTTTAAA	AGG	5.30 ± 1.98	AACCTCACTTA TCTTCTTGT	AGG	24.18 ± 3.22
DYRK1A	21	5'	-13	10.54 ± 0.63	GTATCATTTGAC ATATCTAA	TGG	26.90 ± 1.17	TGTCAAATGAT ACAAACATT	AGG	29.69 ± 0.86
DYRK1A	26	5'	-8	2.33 ± 0.11	CAGCATGGAATG AAAATGAC	CGG	3.33 ± 0.56	CCATGCTGCTG GCCTTCAGA	TGG	20.43 ± 2.40
DYRK1A	29	5'	-5	27.76 ± 0.84	GCAGCATGGAAT GAAAATGA	CGG	17.84 ± 5.46	GCTGCTGGCCT TCAGATGGC	TGG	21.92 ± 3.46
DYRK1A	34	5'	0	10.42 ± 0.90	ATCTGGTCAGAA TATGATAA	AGG	9.13 ± 2.32	TCAGCAACCTC TAACTAACC	AGG	24.14 ± 2.95
DYRK1A	36	5'	2	7.63 ± 0.51	GTGCAAGCCGAA CAGATGAA	AGG	6.65 ± 2.19	TCATTTTCATT CCATGCTGC	TGG	20.61 ± 3.64
DYRK1A	36	5'	2	38.46 ± 0.74	GAACTTACCTGG TTAGTTAG	AGG	20.88 ± 9.09	GGAGTATCAGA AATGACTAT	TGG	30.3 ± 0.7
DYRK1A	41	5'	7	34.41 ± 0.87	GGTCACTGTACT GATGTGAA	TGG	25.68 ± 5.95	GCCAAACATAA GTGACCAAC	AGG	33.1 ± 0.4
DYRK1A	41	5'	7	44.22 ± 0.55	AAAAGACCTAAA CAAAAGAA	TGG	23.20 ± 2.10	TGTGTGAGGAT AAAAGAGTT	GGG	29.4 ± 2.7
DYRK1A	42	5'	8	31.76 ± 1.00	CCGGCCAAGACC TTGAAGCC	AGG	32.50 ± 0.50	CTGGTTGTAGG ATTTGAGTT	AGG	26.7 ± 2.9
DYRK1A	43	5'	9	38.36 ± 0.32	TCACTGTACTGA TGTGAATG	GGG	24.68 ± 4.58	GCCAAACATAA GTGACCAAC	AGG	29.46 ± 3.30
DYRK1A	43	5'	9	28.97 ± 0.32	GTTCCTTAAATA AGAACTTT	AGG	23.60 ± 2.56	TGTCAAATGAT ACAAACATT	AGG	22.4 ± 1.6
DYRK1A	43	5'	9	42.54 ± 1.39	TCAGAGCTTCCT GACACCCA	TGG	14.20 ± 1.50	AATACCTAGTT ACAGGCATT	TGG	24.8 ± 1.0
DYRK1A	47	5'	13	11.90 ± 1.65	TCCTACAAGAAG ATAAGTGA	AGG	6.57 ± 1.36	CATGCAAACCT TCATCTGTT	CGG	30.42 ± 1.14
DYRK1A	47	5'	13	34.45 ± 0.45	TATTACAGAATG AGAGACTG	TGG	30.90 ± 1.40	TTATTTCTGAA GAATATTAA	AGG	27.6 ± 2.5
DYRK1A	78	5'	44	6.63 ± 0.27	CATCTGAAGGCC AGCAGCAT	TGG	10.02 ± 1.17	GCCAAACATAA GTGACCAAC	AGG	22.92 ± 5.16

(Supp Table 2 Continued)

DYRK1A	87	5'	53	N.D.	GAAGATAAGTGA GGTTTAAA	AGG	2.90 ± 0.82	TCATTTTCATT CCATGCTGC	TGG	17.30 ± 1.62
DYRK1A	98	5'	64	N.D.	GAAGATAAGTGA GGTTTAAA	AGG	2.16 ± 0.48	CCATGCTGCTG GCCTTCAGA	TGG	24.75 ± 2.50
DYRK1A	132	5'	98	N.D.	TATCATTTGACA TATCTAAT	TGG	8.21 ± 2.83	TCATTTTCATT CCATGCTGC	TGG	14.61 ± 4.10
DYRK1A	156	5'	122	N.D.	TCCTACAAGAAG ATAAGTGA	AGG	9.99 ± 4.12	GCCAAACATAA GTGACCAAC	AGG	19.74 ± 2.91
DYRK1A	192	5'	158	N.D.	AACTTTTCTAAC TACAAACA	AGG	5.74 ± 2.24	TCATTTTCATT CCATGCTGC	TGG	21.37
GRIN2B	164	3'	-198	N.D.	CCAACACCAACC AGAACTTG	GGG	2.95 ± 0.21	CTGGTAGATGG AGTTGGGTT	TGG	17.25 ± 1.30
GRIN2B	66	3'	-100	N.D.	ACAGCAATGCCA ATGCTGGG	GGG	18.00 ± 2.31	AGTGCTGTTCT CCCAAGTTC	TGG	28.64 ± 0.69
GRIN2B	41	3'	-75	N.D.	GTGGAAATCATC TTTCTCGT	TGG	14.56 ± 7.84	GGCATTGCTGT CATCCTCGT	GGG	21.26 ± 2.68
GRIN2B	15	3'	-49	N.D.	TCTGCTGCCTGA CACGGCCA	AGG	4.24 ± 0.79	TCCCAAGTTCT GGTTGGTGT	TGG	19.64 ± 0.23
GRIN2B	1	3'	-35	N.D.	CGAGCTCTGCTG CCTGACAC	CGG	2.99 ± 0.31	TTGGCCGTCCT GGCCGTGTC	AGG	4.74 ± 0.15
GRIN2B	10	5'	-24	1.04 ± 0.53	TCCTTGATGGCC ACCTCGTC	CGG	2.25 ± 1.08	TTCCGACGAGG TGGCCATCA	AGG	17.13 ± 2.90
GRIN2B	19	5'	-15	5.93 ± 1.25	ATGACAGCAATG CCAATGCT	TGG	16.46 ± 2.28	TGGCATTGCTG TCATCCTCG	TGG	16.35 ± 1.25
GRIN2B	24	5'	-10	2.28 ± 0.34	AGCAATGCCAAT GCTGGGGG	GGG	3.19 ± 0.51	TGGCATTGCTG TCATCCTCG	TGG	15.17 ± 2.02
GRIN2B	29	5'	-5	1.45 ± 0.12	GCCAACACCAAC CAGAACTT	TGG	17.80 ± 2.30	TTGGCCGTCCT GGCCGTGTC	AGG	4.46 ± 1.35
GRIN2B	31	5'	-3	11.80 ± 0.29	GGAGAACAGCAC TCCGCTCT	TGG	21.80 ± 1.40	TCCCAAGTTCT GGTTGGTGT	TGG	21.33 ± 0.63
GRIN2B	34	5'	0	24.24 ± 0.23	ATGACAGCAATG CCAATGCT	TGG	19.48 ± 1.88	CCTCGTGGGCA CTTCCGACG	AGG	21.19 ± 3.42
GRIN2B	35	5'	1	20.83 ± 0.95	TGACAGCAATGC CAATGCTG	GGG	21.44 ± 3.02	CCTCGTGGGCA CTTCCGACG	AGG	24.11 ± 0.14
GRIN2B	39	5'	5	9.60 ± 0.25	AGCAATGCCAAT GCTGGGGG	GGG	4.19 ± 0.58	CCTCGTGGGCA CTTCCGACG	AGG	21.78 ± 1.70
GRIN2B	46	5'	12	18.96 ± 0.93	ATGACAGCAATG CCAATGCT	TGG	20.45 ± 0.98	TTCCGACGAGG TGGCCATCA	AGG	13.21 ± 0.74
GRIN2B	51	5'	17	5.33 ± 0.57	AGCAATGCCAAT GCTGGGGG	GGG	4.93 ± 2.06	TTCCGACGAGG TGGCCATCA	AGG	12.51 ± 1.21
GRIN2B	90	5'	56	7.31 ± 0.83	GAGAACAGCACT CCGCTCTG	GGG	3.09 ± 0.54	CAGAAGAGCCC CCCCAGCAT	TGG	25.02 ± 1.86
GRIN2B	106	5'	72	10.56 ± 1.21	GCCAACACCAAC CAGAACTT	TGG	25.29 ± 1.65	CGTGGGCACTT CCGACGAGG	TGG	23.32 ± 0.78
GRIN2B	133	5'	99	2.66 ± 0.89	CTGCCTGACACG GCCAGGAC	CGG	4.34 ± 0.62	TGATTTCCACC ATCTCTCCG	TGG	20.95 ± 0.79
GRIN2B	176	5'	142	N.D.	GAGAACAGCACT CCGCTCTG	GGG	2.96 ± 0.93	TGATTTCCACC ATCTCTCCG	TGG	19.77 ± 2.20
GRIN2B	232	5'	198	N.D.	CTGCCTGACACG GCCAGGAC	CGG	6.17 ± 2.09	TGACCGGAAGA TCCAGGGGG	TGG	23.36 ± 2.34

(Supp Table 2 Continued)

MeCP2	38	5'	4	12.16 ± 2.58	GTCCAACCTTCA GGCAAGGT	GGG	24.11 ± 3.48	AAGCTTAAACA AAGGAAGTC	TGG	35.76 ± 2.65
MeCP2	34	5'	0	11.72 ± 2.40	GCGCTGTTTGGG GGAAGCCG	AGG	N.T	GGCTCCATTAT CCGTGACCG	GGG	N.T
VEGFA	50	5'	16	23.92 ± 0.54	GGGTGGGGGGAG TTTGCTCC	TGG	14.15 ± 1.07	TCCCTCTTTAG CCAGAGCCG	GGG	33.59 ± 0.88
VEGFA	54	5'	20	16.32 ± 1.01	GACCCCCTCCAC CCCGCCTC	CGG	24.52 ± 1.80	GAAACTTTTCG TCCAACTTC	TGG	10.45 ± 1.14

Supplementary Table 3 List of sgRNAs used in the study

Gene	sgRNA ID	Guide sequence (5′ to 3′)	PAM	Strand	Species
EMX1	1	GAGTCCGAGCAGAAGAAGAA	GGG	+	*H. sapiens*
EMX1	2	GGAAGGGCCTGAGTCCGAGCAGAAGAAGAA	GGG	+	*H. sapiens*
EMX1	3	GGCCTCCAAGGAGTCCGAGCAGAAGAAGAA	GGG	+	*H. sapiens*
EMX1	4	AGGCCCCAGTGGCTGCTCTG	GGG	+	*H. sapiens*
EMX1	5	GGGGCACAGATGAGAAACTC	AGG	—	*H. sapiens*
EMX1	6	TCACCTGGGCCAGGGAGGGA	GGG	—	*H. sapiens*
EMX1	7	TGAAGGTGTGGTTCCAGAAC	CGG	+	*H. sapiens*
EMX1	8	AGGTGTGGTTCCAGAACCGG	AGG	+	*H. sapiens*
EMX1	9	GCCGTTTGTACTTTGTCCTC	CGG	—	*H. sapiens*
EMX1	10	CAAACGGCAGAAGCTGGAGG	AGG	+	*H. sapiens*
EMX1	11	CGGCAGAAGCTGGAGGAGGA	AGG	+	*H. sapiens*
EMX1	12	TGAGTCCGAGCAGAAGAAGA	AGG	+	*H. sapiens*
EMX1	13	GGAGCCCTTCTTCTTCTGCT	CGG	—	*H. sapiens*
EMX1	14	AGGGCTCCCATCACATCAAC	CGG	+	*H. sapiens*
EMX1	15	TGCGCCACCGGTTGATGTGA	TGG	—	*H. sapiens*
EMX1	16	TTGCCACGAAGCAGGCCAAT	GGG	+	*H. sapiens*
EMX1	17	CACGAAGCAGGCCAATGGGG	AGG	+	*H. sapiens*
EMX1	18	TCACCTCCAATGACTAGGGT	GGG	+	*H. sapiens*
EMX1	19	GGCAGAGTGCTGCTTGCTGC	TGG	+	*H. sapiens*
EMX1	20	GACATCGATGTCCTCCCCAT	TGG	—	*H. sapiens*
EMX1	21	GTCACCTCCAATGACTAGGG	TGG	+	*H. sapiens*
EMX1	22	GGGCAACCACAAACCCACGA	GGG	+	*H. sapiens*
EMX1	23	ACTCTGCCCTCGTGGGTTTG	TGG	—	*H. sapiens*
EMX1	24	CAAGCAGCACTCTGCCCTCG	TGG	—	*H. sapiens*
EMX1	25	TTCTTCTTCTGCTCGGACTC	AGG	—	*H. sapiens*
EMX1	26	CTCCCCATTGGCCTGCTTCG	AGG	—	*H. sapiens*
EMX1	27	GTCACCTCCAATGACTAGGG	TGG	+	*H. sapiens*
DYRK1A	28	GAACTTACCTGGTTAGTTAG	AGG	—	*H. sapiens*
DYRK1A	29	GGAGTATCAGAAATGACTAT	TGG	+	*H. sapiens*
DYRK1A	30	GGTCACTGTACTGATGTGAA	TGG	—	*H. sapiens*
DYRK1A	31	GCCAAACATAAGTGACCAAC	AGG	+	*H. sapiens*
DYRK1A	32	GTTCCTTAAATAAGAACTTT	AGG	—	*H. sapiens*
DYRK1A	33	TGTCAAATGATACAAACATT	AGG	+	*H. sapiens*
DYRK1A	34	ATCTGGTCAGAATATGATAA	AGG	—	*H. sapiens*
DYRK1A	35	GTCACTGTACTGATGTGAAT	TGG	—	*H. sapiens*
DYRK1A	36	CATCTGAAGGCCAGCAGCAT	TGG	—	*H. sapiens*
DYRK1A	37	TGATAAGGCAGAAACCTGTT	TGG	—	*H. sapiens*
DYRK1A	38	GAAGATAAGTGAGGTTTAAA	AGG	—	*H. sapiens*
DYRK1A	39	GTATCATTTGACATATCTAA	TGG	—	*H. sapiens*
DYRK1A	40	CAGCATGGAATGAAAATGAC	CGG	—	*H. sapiens*
DYRK1A	41	GCAGCATGGAATGAAAATGA	CGG	—	*H. sapiens*
DYRK1A	42	GTGCAAGCCGAACAGATGAA	AGG	—	*H. sapiens*
DYRK1A	43	TCACTGTACTGATGTGAATG	GGG	—	*H. sapiens*
DYRK1A	44	TCCTACAAGAAGATAAGTGA	AGG	—	*H. sapiens*
DYRK1A	45	TATCATTTGACATATCTAAT	TGG	—	*H. sapiens*
DYRK1A	46	AACTTTTCTAACTACAAACA	AGG	—	*H. sapiens*
DYRK1A	47	AACCTCACTTATCTTCTTGT	AGG	+	*H. sapiens*
DYRK1A	48	CTCACTTATCTTCTTGTAGG	AGG	+	*H. sapiens*

(Supp Table 3 Continued)

DYRK1A	49	CCATGCTGCTGGCCTTCAGA	TGG	+	*H. sapiens*
DYRK1A	50	GCTGCTGGCCTTCAGATGGC	TGG	+	*H. sapiens*
DYRK1A	51	TCAGCAACCTCTAACTAACC	AGG	+	*H. sapiens*
DYRK1A	52	TCATTTTCATTCCATGCTGC	TGG	+	*H. sapiens*
DYRK1A	53	CATGCAAACCTTCATCTGTT	CGG	+	*H. sapiens*
DYRK1A	54	TATTACAGAATGAGAGACTG	TGG	—	*H. sapiens*
DYRK1A	55	TTATTTCTGAAGAATATTAA	AGG	+	*H. sapiens*
DYRK1A	56	AAAAGACCTAAACAAAAGAA	TGG	—	*H. sapiens*
DYRK1A	57	TGTGTGAGGATAAAAGAGTT	GGG	+	*H. sapiens*
DYRK1A	58	CCGGCCAAGACCTTGAAGCC	AGG	—	*H. sapiens*
DYRK1A	59	CTGGTTGTAGGATTTGAGTT	AGG	+	*H. sapiens*
DYRK1A	60	TCAGAGCTTCCTGACACCCA	TGG	—	*H. sapiens*
DYRK1A	61	AATACCTAGTTACAGGCATT	TGG	+	*H. sapiens*
GRIN2B	62	GGTGATGATGCTCTTTGGGT	CGG	—	*H. sapiens*
GRIN2B	63	TCTGTGATCTCATGTCTGAC	CGG	+	*H. sapiens*
GRIN2B	64	CAGCAATGCCAATGCTGGGG	GGG	—	*H. sapiens*
GRIN2B	65	CCTCGTGGGCACTTCCGACG	AGG	+	*H. sapiens*
GRIN2B	66	TTTCTCGTGGGCATCCTTGA	TGG	—	*H. sapiens*
GRIN2B	67	TGATTTCCACCATCTCTCCG	TGG	+	*H. sapiens*
GRIN2B	68	GGAGAACAGCACTCCGCTCT	GGG	—	*H. sapiens*
GRIN2B	69	CTGGTTGGTGTTGGCCGTCC	TGG	+	*H. sapiens*
GRIN2B	70	CCAACACCAACCAGAACTTG	GGG	—	*H. sapiens*
GRIN2B	71	ACAGCAATGCCAATGCTGGG	GGG	—	*H. sapiens*
GRIN2B	72	GTGGAAATCATCTTTCTCGT	TGG	—	*H. sapiens*
GRIN2B	73	TCTGCTGCCTGACACGGCCA	AGG	—	*H. sapiens*
GRIN2B	74	CGAGCTCTGCTGCCTGACAC	CGG	—	*H. sapiens*
GRIN2B	75	TCCTTGATGGCCACCTCGTC	CGG	—	*H. sapiens*
GRIN2B	76	ATGACAGCAATGCCAATGCT	TGG	—	*H. sapiens*
GRIN2B	77	AGCAATGCCAATGCTGGGGG	GGG	—	*H. sapiens*
GRIN2B	78	GCCAACACCAACCAGAACTT	TGG	—	*H. sapiens*
GRIN2B	79	TGACAGCAATGCCAATGCTG	GGG	—	*H. sapiens*
GRIN2B	80	GAGAACAGCACTCCGCTCTG	GGG	—	*H. sapiens*
GRIN2B	81	CTGCCTGACACGGCCAGGAC	CGG	—	*H. sapiens*
GRIN2B	82	CTGGTAGATGGAGTTGGGTT	TGG	+	*H. sapiens*
GRIN2B	83	AGTGCTGTTCTCCCAAGTTC	TGG	+	*H. sapiens*
GRIN2B	84	GGCATTGCTGTCATCCTCGT	GGG	+	*H. sapiens*
GRIN2B	85	TCCCAAGTTCTGGTTGGTGT	TGG	+	*H. sapiens*
GRIN2B	86	TTGGCCGTCCTGGCCGTGTC	AGG	+	*H. sapiens*
GRIN2B	87	TTCCGACGAGGTGGCCATCA	AGG	+	*H. sapiens*
GRIN2B	88	TGGCATTGCTGTCATCCTCG	TGG	+	*H. sapiens*
GRIN2B	89	CAGAAGAGCCCCCCCAGCAT	TGG	+	*H. sapiens*
GRIN2B	90	CGTGGGCACTTCCGACGAGG	TGG	+	*H. sapiens*
GRIN2B	91	TGACCGGAAGATCCAGGGGG	TGG	+	*H. sapiens*
MeCP2	92	GTCCAACCTTCAGGCAAGGT	GGG	—	*M. musculus*
MeCP2	93	AAGCTTAAACAAAGGAAGTC	TGG	+	*M. musculus*
MeCP2	94	GCGCTGTTTGGGGGAAGCCG	AGG	—	*M. musculus*
MeCP2	95	GGCTCCATTATCCGTGACCG	GGG	+	*M. musculus*
VEGFA	96	GGGTGGGGGGAGTTTGCTCC	TGG	—	*H. sapiens*
VEGFA	97	TCCCTCTTTAGCCAGAGCCG	GGG	+	*H. sapiens*
VEGFA	98	GACCCCCTCCACCCCGCCTC	CGG	—	*H. sapiens*
VEGFA	99	GAAACTTTTCGTCCAACTTC	TGG	+	*H. sapiens*

Supplementary Table 4 Data collection and refinement statistics

	Native Cas9	SeMet Cas9
Data collection		
Beamline	SPring-8 BL32XU	SPring-8 BL41XU
Wavelength (Å)	1.0000	0.97911
Space group	*P*1	*P*1
Cell dimensions		
a, *b*, *c* (Å)	76.7, 105.7, 126.8	76.2, 104.5, 125.5
α, β, γ (°)	97.7, 98.4, 100.3	97.0, 98.2, 101.1
Resolution (Å)	50–2.5 (2.65–2.5)	50–2.6 (2.67–2.6)
R_{sym}	0.06 (1.06)	0.17 (1.96)
$I/\sigma I$	25.3 (2.1)	12.6 (1.4)
Completeness (%)	98.2 (96.2)	99.9 (99.9)
Redundancy	7.9 (7.9)	19.1 (15.9)
CC(1/2)	1.00 (0.795)	1.00 (0.736)
Refinement		
Resolution (Å)	50–2.5	
No. reflections	130,396	
R_{work}/ R_{free}	0.222 / 0.253	
No. atoms		
Protein	18,997	
Nucleic acid	5,012	
Solvent	144	
B-factors (Å^2)		
Protein	80.2	
Nucleic acid	82.4	
Solvent	51.4	
R.m.s. deviations		
Bond lengths (Å)	0.002	
Bond angles (°)	0.526	
Ramachandran plot		
Favored region	97.0%	
Allowed region	3.0%	
Outlier region	0.0%	

*Highest resolution shell is shown in parentheses.

Supplementary Table 5 List of sgRNA pairs used with Cas9 nickases

		sgRNA 1 target site		**sgRNA 2 target site**	
Gene	Offset length (bp)	Guide sequence (5' to 3')	PAM	Guide sequence (5' to 3')	PAM
EMX1	20	GGGGCACAGATGAGAAACTC	AGG	AGGTGTGGTTCCAGAACCGG	AGG
EMX1	17	GGGGCACAGATGAGAAACTC	AGG	TGAAGGTGTGGTTCCAGAAC	CGG
EMX1	4	TGCGCCACCGGTTGATGTGA	TGG	CACGAAGCAGGCCAATGGGG	AGG
EMX1	0	TCACCTGGGCCAGGGAGGGA	GGG	AGGTGTGGTTCCAGAACCGG	AGG
EMX1	-29	GGGCAACCACAAACCCACGA	GGG	ACTCTGCCCTCGTGGGTTTG	TGG

Supplementary Table 6 List of Cas9 orthologs and predicted RNA components

Cas9	class	direct repeat	tracrRNA	sgRNA
P. lavamentivorans	IIC	GCUGCGGAU UGCGGCCGU CUCUCGAUU UGCUACUCU	UAGCAAAUCGAGAGGCGGUCGCUU UUCGCAAGCAAAUUGACCCCUUGU GCGGGCUCGGCAUCCCAAGGUCAG CUGCCGGUUAUUAUCGAAAAGGCC CACCGCAAGCAGCGCGUGGGCCUU UUU	GCUGCGGAUUGCGGGAA AUCGCUUUUCGCAAGCA AAUUGACCCCUUGUGCG GGCUCGGCAUCCCAAGG UCAGCUGCCGGUUAUUA UCGAAAAGGCCCACCGC AAGCAGCGCGUGGGCCU UUU
C. diphtheriae	IIC	ACUGGGGUU CAGUUCUCA AAAACCCUG AUAGACUUC	AGUCACUAACUUAAUUAAAUAGAA CUGAACCUCAGUAAGCAUUGGCUC GUUUCCAAUGUUGAUUGCUCCGCC GGUGCUCCUUAUUUUUAAGGGCGC CGGCUUUCUU	ACUGGGGUUCAGGAAAC UGAACCUCAGUAAGCAU UGGCUCGUUUCCAAUGU UGAUUGCUCCGCCGGUG CUCCUUAUUUUUAAGGG CGCCGGCUUUU
S. pasteurianus	IIA	GUUUUUGUA CUCUCAAGA UUUAAGUAA CCGUAAAAC	CUUGCACGGUUACUUAAAUCUUGC UGAGCCUACAAAGAUAAGGCUUUA UGCCGAAUUCAAGCACCCCAUGUU UUGACAUGAGGUGCUUUU	GUUUUUGUACUCGAAAG AGCCUACAAAGAUAAGG CUUUAUGCCGAAUUCAA GCACCCCAUGUUUUGAC AUGAGGUGCUUUU
N. cinerea	IIC	GUUGUAGCU CCCAUUCUC AUUUCGCAG UGCUACAAU	AUUGUCGCACUGCGAAAUGAGAAC CGUUGCUACAAUAAGGCCGUCUGA AAAGAUGUGCCGCAACGCUCUGCC CCUUAAAGCUUCUGCUUUAAGGGG CAUCGUUUAUUUCGGUUAAAAAUG CCGUCUGAAACCGGUUUUU	GUUGUAGCUCCCAUUCU CGAAAGAGAACCGUUGC UACAAUAAGGCCGUCUG AAAAGAUGUGCCGCAAC GCUCUGCCCCUUAAAGC UUCUGCUUUAAGGGGCA UCGUUUAUUUCGGUUAA AAAUGCCGUCUGAAACC GGUUUUUAGGUUUCAGA CGGCAUUUU
S. aureus	IIA	GUUUUAGUA CUCUGUAAU UUUAGGUAU GAGGUAGAC	AUUGUACUUAUACCUAAAAUUACA GAAUCUACUAAAACAAGGCAAAAU GCCGUGUUUAUCUCGUCAACUUGU UGGCGAGAUUUUU	GUUUUAGUACUCUGGAA ACAGAAUCUACUAAAAC AAGGCAAAAUGCCGUGU UUAUCUCGUCAACUUGU UGGCGAGAUUUU
C. lari	IIC	GUUUUAGUC UCUUUUUAA AUUUCUUUA UGAUAAAAU	AAUUCUUGCUAAAGAAAUUUAAAA AGAGACUAAAAUAAGUGGUUUUUG GUCAUCCACGCAGGGUUACAAUCC CUUUAAAACCAUUAAAAUUCAAAU AAACUAGGUUGUAUCAACUUAGUU UUUU	GUUUUAGUCUCUGAAAA GAGACUAAAAUAAGUGG UUUUUGGUCAUCCACGC AGGGUUACAAUCCCUUU AAAACCAUUAAAAUUCA AAUAAACUAGGUUGUAU CAACUUAGUUUU
S. pyogenes	IIA	GUUUUAGAG CUAUGCUGU UUUGAAUGG UCCCAAAAC	GUUGGAACCAUUCAAAACAGCAUA GCAAGUUAAAAUAAGGCUAGUCCG UUAUCAACUUGAAAAAGUGGCACC GAGUCGGUGCUUUUU	GUUUUAGAGCUAGAAAU AGCAAGUUAAAAUAAGG CUAGUCCGUUAUCAACU UGAAAAAGUGGCACCGA GUCGGUGCUUUU
S. thermophilus	IIA	GUUUUUGUA CUCUCAAGA UUUAAGUAA CUGUACAAC	CUUACACAGUUACUUAAAUCUUGC AGAAGCUACAAAGAUAAGGCUUCA UGCCGAAAUCAACACCCUGUCAUU UUAUGGCAGGGUGUUUU	GUUUUUGUACUCGAAAG AAGCUACAAAGAUAAGG CUUCAUGCCGAAAUCAA CACCCUGUCAUUUUAUG GCAGGGUGUUUU

Supplementary Table 7 Targets used for in vitro (cell lysate) PAM validation

Cas9	**Consensus**	***in vitro* lysate targets (Dyrk1a)**	**PAM**	***Gene (PCR amplicon)***
P. lavamentivorans	NNNCATN	TAATCACTATGGATCTTCTA	TACCATT	*DYRK1A*
P. lavamentivorans	NNNCATN	TCTTGTAGGAGGAGAGACTT	CAGCATG	*DYRK1A*
C. diphtheriae	NGGNNNN	GGTGCAAGCCGAACAGATGA	TGGACAG	*DYRK1A*
C. diphtheriae	NGGNNNN	TATCCTAAAGTTCTTATTTA	AGGTTTG	*DYRK1A*
S. pasteurianus	NNGTGAN	TTAATTTATGAAAATCTCGT	AGGTGAA	*DYRK1A*
S. pasteurianus	NNGTGAN	ATGCCCCATTCACATCAGTA	CAGTGAC	*DYRK1A*
N. cinerea	NNNNGAT	GTGTTGAGTAACATATACCT	GTTTGTA	*DYRK1A*
N. cinerea	NNNNGAT	TAACTAACCAGGTAAGTTCA	TGGAGTA	*DYRK1A*
S. aureus	NNGRRNN	AATGATACAAACATTAGGAT	ATGAATA	*DYRK1A*
S. aureus	NNGRRNN	ATGTCAAATGATACAAACAT	TAGGATA	*DYRK1A*
C. lari	NNGGGNN	GGTCACTGTACTGATGTGAA	TGGGGCA	*DYRK1A*
C. lari	NNGGGNN	CGGTCACTGTACTGATGTGA	ATGGGGC	*DYRK1A*
S. pyogenes	NGGNNNN	TGTCAAATGATACAAACATT	AGGATAT	*DYRK1A*
S. pyogenes	NGGNNNN	AACCTCACTTATCTTCTTGT	AGGAGGA	*DYRK1A*
S. thermophilus	NNAGAAW	CCAGGTAAGTTCATGGAGTA	TCAGAAA	*DYRK1A*
S. thermophilus	NNAGAAW	TAACATATACCTGTTTGTAG	TTAGAAA	*DYRK1A*

Supplementary Table 8 Targets used for testing ortholog activity in human cells

Cas9	Consensus	Targets	PAM	Gene	Cell type	indel (%)
C. diphtheriae	NGGNNNN	TCACCTCCAATGACTAGGGT	GGGCAAC	*EMX1*	293FT	N.D.
C. diphtheriae	NGGNNNN	TGACGGTGCAAGCCGAACAGATGA	TGGACAG	*DYRK1A*	293FT	N.D.
C. diphtheriae	NGGNNNN	ACCTGGTGGGCGACGTGCTG	GGGAGTC	*DYRK1A*	293FT	N.D.
C. diphtheriae	NGGNNNN	ATGGAGCAGTCTCAGTCTTC	GGGCACC	*DYRK1A*	293FT	N.D.
N. cinerea	NNNNGAT	GAATGAAAATGACGGTGCAAGCCG	AACAGAT	*DYRK1A*	293FT	N.D.
N. cinerea	NNNNGAT	TTAATGGTATAGAAGATCCA	TAGTGAT	*DYRK1A*	293FT	N.D.
C. lari	NNGGGNN	TGTCACCTCCAATGACTAGG	GTGGGCA	*EMX1*	293FT	N.D.
C. lari	NNGGGNN	CCATGGAGCAGTCTCAGTCT	TCGGGCA	*DYRK1A*	293FT	N.D.
C. lari	NNGGGNN	GCACCAGCATCGGCACAGTG	GTGGGCA	*DYRK1A*	293FT	N.D.
C. lari	NNGGGNN	CGACGGTCACTGTACTGATGTGAA	TGGGGCA	*DYRK1A*	293FT	N.D.
P. lavamentivorans	NNNCATN	CCGAGCAGAAGAAGAAGGGC	TCCCATC	*EMX1*	293FT	N.D.
P. lavamentivorans	NNNCATN	ATTTTAATCACTATGGATCTTCTA	TACCATT	*DYRK1A*	293FT	N.D.
P. lavamentivorans	NNNCATN	CCAAAACTCGAATTCAACCT	GGTCATA	*DYRK1A*	293FT	N.D.
P. lavamentivorans	NNNCATN	TGCAGCACAGTTTCTTCAAG	GAGCATA	*DYRK1A*	293FT	N.D.
S. pasteurianus	NNGTGAN	GTTCTTAATTTATGAAAATCTCGT	AGGTGAA	*DYRK1A*	293FT	N.D.
S. pyogenes	NGGNNNN	GAGTCCGAGCAGAAGAAGAA	GGGCTCC	*EMX1*	293FT	33.3
S. pyogenes	NGGNNNN	TGACGGTGCAAGCCGAACAGATGA	TGGACAG	*DYRK1A*	293FT	3.0
S. pyogenes	NGGNNNN	ATCAGAAAAGAAAGAACAGC	TGGAGTC	*Sqle*	Hepa1-6	14.5
S. pyogenes	NGGNNNN	GCAACAACAAGATCTGTGGC	TGGAATT	*HmgCR*	Hepa1-6	13.5
S. pyogenes	NGGNNNN	TGTTCCCACAATAACTTCCC	AGGGGTG	*HmgCR*	Hepa1-6	11.6
S. thermophilus	NNAGAAW	TGAGTAACATATACCTGTTTGTAG	TTAGAAA	*DYRK1A*	293FT	5.0
S. aureus	NNGRRNN	CAACCACAAACCCACGAGGG	CAGAGTG	*EMX1*	293FT	15.9
S. aureus	NNGRRNN	TAGGGTTAGGGGCCCCAGGC	CGGGGTC	*EMX1*	*293FT*	13.0
S. aureus	NNGRRNN	CCTCTAACTAACCAGGTAAGTTCA	TGGAGTA	*DYRK1A*	293FT	6.7
S. aureus	NNGRRNN	TAAGAGAGTAGGCTGGTAGA	TGGAGTT	*GRIN2B*	293FT	24.2
S. aureus	NNGRRNN	GAGTAGGCTGGTAGATGGAG	TTGGGTT	*GRIN2B*	293FT	31.7
S. aureus	NNGRRNN	GTTGAAGATGAAGCCCAGAG	CGGAGTG	*GRIN2B*	293FT	13.4
S. aureus	NNGRRNN	TGGATGCCCAGGATGGGGGT	GAGAGTA	*GRIN2B*	293FT	18.7
S. aureus	NNGRRNN	AAAGAAAGAGCATGTTAAAA	TAGGATA	*GRIN2B*	293FT	N.D.
S. aureus	NNGRRNN	TCAGACATGAGATCACAGAT	GCGGGTG	*GRIN2B*	293FT	29.3
S. aureus	NNGRRNN	GATGCGGGTGATGATGCTCT	TTGGGTC	*GRIN2B*	293FT	17.6
S. aureus	NNGRRNN	TCATGGCTACCAGTTCCACC	CGGGGTA	*GRIN2B*	293FT	26.6
S. aureus	NNGRRNN	CCCGGGTGGAACTGGTAGCC	ATGAATG	*GRIN2B*	293FT	26.2
S. aureus	NNGRRNN	CTTCCGACGAGGTGGCCATC	AAGGATT	*GRIN2B*	293FT	7.6
S. aureus	NNGRRNN	CACCATCTCTCCGTGGTACC	CCGGGTG	*GRIN2B*	293FT	18.2
S. aureus	NNGRRNN	ATCTCTTAGATACCAGCATC	CAGGGTG	*Pcsk9*	Hepa1-6	4.6
S. aureus	NNGRRNN	TCAATCTCCCGATGGGCACC	CTGGATG	*Pcsk9*	Hepa1-6	2.6
S. aureus	NNGRRNN	GCCCATCGGGAGATTGAGGG	CAGGGTC	*Pcsk9*	Hepa1-6	9.7
S. aureus	NNGRRNN	ACTTCAACAGCGTGCCGGAG	GAGGATG	*Pcsk9*	Hepa1-6	6.2
S. aureus	NNGRRNN	CCGCTGACCACACCTGCCAG	GTGGGTG	*Pcsk9*	Hepa1-6	8.3
S. aureus	NNGRRNN	TGGCAGGTGTGGTCAGCGGC	CGGGATG	*Pcsk9*	Hepa1-6	3.4
S. aureus	NNGRRNN	ATCAGAAAAGAAAGAACAGC	TGGAGTC	*Sqle*	Hepa1-6	21.1
S. aureus	NNGRRNN	GCAACAACAAGATCTGTGGC	TGGAATT	*HmgCR*	Hepa1-6	7.1
S. aureus	NNGRRNN	TGTTCCCACAATAACTTCCC	AGGGGTG	*HmgCR*	Hepa1-6	9.5

Supplementary Table 9 Targets used for SaCas9 PAM determination in mammalian cells

Cas9	Targets	PAM	Gene	Cell type	indel (%)
S. aureus	GAGGACCGCCCTGGGCCTGG	GAGAAT	*Rosa26*	Hepa1-6	9
S. aureus	CACGAGGGGAAGAGGGGGCA	AGGGAT	*Rosa26*	Hepa1-6	12
S. aureus	CGCCCATCTTCTAGAAAGAC	TGGAGT	*Rosa26*	Hepa1-6	16
S. aureus	AGTCTTTCTAGAAGATGGGC	GGGAGT	*Rosa26*	Hepa1-6	14
S. aureus	GTGTGGGCGTTGTCCTGCAG	GGGAAT	*Rosa26*	Hepa1-6	13
S. aureus	TAGGGGCAAATAGGAAAATG	GAGGAT	*Rosa26*	Hepa1-6	0
S. aureus	CAAATAGGAAAATGGAGGAT	AGGAGT	*Rosa26*	Hepa1-6	24
S. aureus	AATGGAGGATAGGAGTCATC	TGGGGT	*Rosa26*	Hepa1-6	17
S. aureus	TCCTCATGGAAATCTCCGAG	GCGGAT	*Rosa26*	Hepa1-6	17
S. aureus	AGGAGATAAAGACATGTCAC	CCGAGT	*Rosa26*	Hepa1-6	0
S. aureus	CTAAGCAGGAGAGTATAAAC	TCGGGT	*Rosa26*	HEK 293FT	0
S. aureus	CTGTAGTAGGATCTAAGCAG	GAGAGT	*Rosa26*	HEK 293FT	0
S. aureus	CACTGTATTTCATACTGTAG	TAGGAT	*Rosa26*	HEK 293FT	0
S. aureus	CTGCAGAAGGAGCGGGAGAA	ATGGAT	*Rosa26*	HEK 293FT	17
S. aureus	GAGTGTTGCAATACCTTTCT	GGGAGT	*Rosa26*	HEK 293FT	17
S. aureus	CCTGGACACCCCGTTCTCCT	GTGGAT	*AAVS1*	HEK 293FT	5
S. aureus	ACAGCATGTTTGCTGCCTCC	AGGGAT	*AAVS1*	HEK 293FT	13
S. aureus	GTGGTCCCAGCTCGGGGACA	CAGGAT	*AAVS1*	HEK 293FT	30
S. aureus	CGGTTAATGTGGCTCTGGTT	CTGGGT	*AAVS1*	HEK 293FT	35
S. aureus	TGTCCCTAGTGGCCCCACTG	TGGGGT	*AAVS1*	HEK 293FT	31
S. aureus	TCCTTCCTAGTCTCCTGATA	TTGGGT	*AAVS1*	HEK 293FT	34
S. aureus	CCTGAAGTGGACATAGGGGC	CCGGGT	*AAVS1*	HEK 293FT	0
S. aureus	GAGAGATGGCTCCAGGAAAT	GGGGGT	*AAVS1*	HEK 293FT	16
S. aureus	TTGCTTACGATGGAGCCAGA	GAGGAT	*AAVS1*	HEK 293FT	0
S. aureus	GAGCCACATTAACCGGCCCT	GGGAAT	*AAVS1*	HEK 293FT	32
S. aureus	CACAGTGGGGCCACTAGGGA	CAGGAT	*AAVS1*	HEK 293FT	27
S. aureus	GACTAGGAAGGAGGAGGCCT	AAGGAT	*AAVS1*	HEK 293FT	23
S. aureus	GAATCTGCCTAACAGGAGGT	GGGGGT	*AAVS1*	HEK 293FT	26
S. aureus	TGGGGGTGTGTCACCAGATA	AGGAAT	*AAVS1*	HEK 293FT	15
S. aureus	CCCTGCCAAGCTCTCCCTCC	CAGGAT	*AAVS1*	HEK 293FT	18
S. aureus	CTGGGAGGGAGAGCTTGGCA	GGGGGT	*AAVS1*	HEK 293FT	0
S. aureus	CAGGGGGTGGGAGGGAAGGG	GGGGAT	*AAVS1*	HEK 293FT	0
S. aureus	GGTGGCTAAAGCCAGGGAGA	CGGGGT	*AAVS1*	HEK 293FT	0
S. aureus	TAGGGTTAGGGGCCCCAGGC	CGGGGT	*EMX1*	HEK 293FT	0
S. aureus	ATGGGAAGACTGAGGCTACA	TAGGGT	*EMX1*	HEK 293FT	0
S. aureus	CATCAGGCTCTCAGCTCAGC	CTGAGT	*EMX1*	HEK 293FT	0
S. aureus	GTGGCTGCTCTGGGGGCCTC	CTGAGT	*EMX1*	HEK 293FT	29
S. aureus	GAAGCTGGAGGAGGAAGGGC	CTGAGT	*EMX1*	HEK 293FT	8
S. aureus	TCGATGTCACCTCCAATGAC	TAGGGT	*EMX1*	HEK 293FT	15
S. aureus	GCAAGCAGCACTCTGCCCTC	GTGGGT	*EMX1*	HEK 293FT	8
S. aureus	CAACCACAAACCCACGAGGG	CAGAGT	*EMX1*	HEK 293FT	32
S. aureus	AAGCCTGGCCAGGGAGTGGC	CAGAGT	*EMX1*	HEK 293FT	7
S. aureus	GCCTCCCCAAAGCCTGGCCA	GGGAGT	*EMX1*	HEK 293FT	28
S. aureus	GGCCAGGCTTTGGGGAGGCC	TGGAGT	*EMX1*	HEK 293FT	24
S. aureus	CAGGCTGAGCTGAGAGCCTG	ATGGGA	*EMX1*	HEK 293FT	9
S. aureus	CTCAACACTCAGGCTGAGCT	GAGAGC	*EMX1*	HEK 293FT	9
S. aureus	GCCTCAACACTCAGGCTGAG	CTGAGA	*EMX1*	HEK 293FT	9
S. aureus	CTGGGGCCTCAACACTCAGG	CTGAGC	*EMX1*	HEK 293FT	8

(Supp Table 9 Continued)

S. aureus	GAGGCCCCCAGAGCAGCCAC	TGGGGC	*EMX1*	HEK 293FT	20
S. aureus	GGAGGCCCCCAGAGCAGCCA	CTGGGG	*EMX1*	HEK 293FT	21
S. aureus	TGAGAAACTCAGGAGGCCCC	CAGAGC	*EMX1*	HEK 293FT	15
S. aureus	GGGGCACAGATGAGAAACTC	AGGAGG	*EMX1*	HEK 293FT	10
S. aureus	AGGGGCACAGATGAGAAACT	CAGGAG	*EMX1*	HEK 293FT	2
S. aureus	AGGGAGGGAGGGGCACAGAT	GAGAAA	*EMX1*	HEK 293FT	5
S. aureus	CCAGGGAGGGAGGGGCACAG	ATGAGA	*EMX1*	HEK 293FT	3
S. aureus	TTCACCTGGGCCAGGGAGGG	AGGGGC	*EMX1*	HEK 293FT	1
S. aureus	CTTCACCTGGGCCAGGGAGG	GAGGGG	*EMX1*	HEK 293FT	8
S. aureus	ACCTTCACCTGGGCCAGGGA	GGGAGG	*EMX1*	HEK 293FT	7
S. aureus	CACCTTCACCTGGGCCAGGG	AGGGAG	*EMX1*	HEK 293FT	6
S. aureus	ACCACACCTTCACCTGGGCC	AGGGAG	*EMX1*	HEK 293FT	5
S. aureus	ACACCTTCACCTGGGCCAGG	GAGGGA	*EMX1*	HEK 293FT	5
S. aureus	CCACACCTTCACCTGGGCCA	GGGAGG	*EMX1*	HEK 293FT	8
S. aureus	AACCACACCTTCACCTGGGC	CAGGGA	*EMX1*	HEK 293FT	6
S. aureus	TTCTGGAACCACACCTTCAC	CTGGGC	*EMX1*	HEK 293FT	7
S. aureus	TGTACTTTGTCCTCCGGTTC	TGGAAC	*EMX1*	HEK 293FT	2
S. aureus	TTGTACTTTGTCCTCCGGTT	CTGGAA	*EMX1*	HEK 293FT	2
S. aureus	GGGAGCCCTTCTTCTTCTGC	TCGGAC	*EMX1*	HEK 293FT	0
S. aureus	GCGCCACCGGTTGATGTGAT	GGGAGC	*EMX1*	HEK 293FT	2
S. aureus	TGCGCCACCGGTTGATGTGA	TGGGAG	*EMX1*	HEK 293FT	7
S. aureus	ATGCGCCACCGGTTGATGTG	ATGGGA	*EMX1*	HEK 293FT	0
S. aureus	CTCTCAGCTCAGCCTGAGTG	TTGAGG	*EMX1*	HEK 293FT	11
S. aureus	TTGAGGCCCCAGTGGCTGCT	CTGGGG	*EMX1*	HEK 293FT	0
S. aureus	TGAGGCCCCAGTGGCTGCTC	TGGGGG	*EMX1*	HEK 293FT	0
S. aureus	GAGGCCCCAGTGGCTGCTCT	GGGGGC	*EMX1*	HEK 293FT	0
S. aureus	CCCCTCCCTCCCTGGCCCAG	GTGAAG	*EMX1*	HEK 293FT	4
S. aureus	CCCAGGTGAAGGTGTGGTTC	CAGAAC	*EMX1*	HEK 293FT	4
S. aureus	GTGAAGGTGTGGTTCCAGAA	CCGGAG	*EMX1*	HEK 293FT	0
S. aureus	TGAAGGTGTGGTTCCAGAAC	CGGAGG	*EMX1*	HEK 293FT	12
S. aureus	AAGGTGTGGTTCCAGAACCG	GAGGAC	*EMX1*	HEK 293FT	10
S. aureus	GGAGGACAAAGTACAAACGG	CAGAAG	*EMX1*	HEK 293FT	3
S. aureus	CAAAGTACAAACGGCAGAAG	CTGGAG	*EMX1*	HEK 293FT	2
S. aureus	AAAGTACAAACGGCAGAAGC	TGGAGG	*EMX1*	HEK 293FT	3
S. aureus	AGTACAAACGGCAGAAGCTG	GAGGAG	*EMX1*	HEK 293FT	3
S. aureus	GTACAAACGGCAGAAGCTGG	AGGAGG	*EMX1*	HEK 293FT	8
S. aureus	ACAAACGGCAGAAGCTGGAG	GAGGAA	*EMX1*	HEK 293FT	3
S. aureus	CAAACGGCAGAAGCTGGAGG	AGGAAG	*EMX1*	HEK 293FT	4
S. aureus	ACGGCAGAAGCTGGAGGAGG	AAGGGC	*EMX1*	HEK 293FT	26
S. aureus	GGAGGAGGAAGGGCCTGAGT	CCGAGC	*EMX1*	HEK 293FT	5
S. aureus	AGGAAGGGCCTGAGTCCGAG	CAGAAG	*EMX1*	HEK 293FT	13
S. aureus	AAGGGCCTGAGTCCGAGCAG	AAGAAG	*EMX1*	HEK 293FT	8
S. aureus	GGCCTGAGTCCGAGCAGAAG	AAGAAG	*EMX1*	HEK 293FT	1
S. aureus	CTGAGTCCGAGCAGAAGAAG	AAGGGC	*EMX1*	HEK 293FT	1
S. aureus	TCAACCGGTGGCGCATTGCC	ACGAAG	*EMX1*	HEK 293FT	7
S. aureus	GGCCACTCCCTGGCCAGGCT	TTGGGG	*EMX1*	HEK 293FT	0
S. aureus	GCCACTCCCTGGCCAGGCTT	TGGGGA	*EMX1*	HEK 293FT	0
S. aureus	CCACTCCCTGGCCAGGCTTT	GGGGAG	*EMX1*	HEK 293FT	5
S. aureus	CACTCCCTGGCCAGGCTTTG	GGGAGG	*EMX1*	HEK 293FT	7
S. aureus	TGGCCAGGCTTTGGGGAGGC	CTGGAG	*EMX1*	HEK 293FT	0

(Supp Table 9 Continued)

S. aureus	GGCCTCCCCAAAGCCTGGCC	AGGGAG	*EMX1*	HEK 293FT	5
S. aureus	AGGCCTCCCCAAAGCCTGGC	CAGGGA	*EMX1*	HEK 293FT	9
S. aureus	TGTCACCTCCAATGACTAGG	GTGGGC	*EMX1*	HEK 293FT	1
S. aureus	GTGGGCAACCACAAACCCAC	GAGGGC	*EMX1*	HEK 293FT	5
S. aureus	TGGTTGCCCACCCTAGTCAT	TGGAGG	*EMX1*	HEK 293FT	1
S. aureus	GTGGTTGCCCACCCTAGTCA	TTGGAG	*EMX1*	HEK 293FT	1
S. aureus	GGCCTGGAGTCATGGCCCCA	CAGGGC	*EMX1*	HEK 293FT	5
S. aureus	GAGTCATGGCCCCACAGGGC	TTGAAG	*EMX1*	HEK 293FT	7
S. aureus	GCCCCGGGCTTCAAGCCCTG	TGGGGC	*EMX1*	HEK 293FT	0
S. aureus	GGCCCCGGGCTTCAAGCCCT	GTGGGG	*EMX1*	HEK 293FT	3
S. aureus	CATTGCCACGAAGCAGGCCA	ATGGGG	*EMX1*	HEK 293FT	16
S. aureus	ATTGCCACGAAGCAGGCCAA	TGGGGA	*EMX1*	HEK 293FT	10
S. aureus	TTGCCACGAAGCAGGCCAAT	GGGGAG	*EMX1*	HEK 293FT	0
S. aureus	TGCCACGAAGCAGGCCAATG	GGGAGG	*EMX1*	HEK 293FT	15
S. aureus	CCACGAAGCAGGCCAATGGG	GAGGAC	*EMX1*	HEK 293FT	30
S. aureus	GGGTGGGCAACCACAAACCC	ACGAGG	*EMX1*	HEK 293FT	6
S. aureus	GCTGCTGGCCAGGCCCCTGC	GTGGGC	*EMX1*	HEK 293FT	3
S. aureus	GAGTCCAGCTTGGGCCCACG	CAGGGG	*EMX1*	HEK 293FT	6

Supplementary Table 10 Predicted GWOTs for SaCas9 and SpCas9 specificity analysis

Target #	Genome-wide off target site	PAM	Mismatches	Forward priming site	Reverse priming site	SpCas9 indel (%)	SaCas9 indel (%)
On-target	TAGGGTTAGGGGCCCCAGGC	CGGGG	NA	CACTGTGTCCTCTTCCTGCC	ATGAGAAACTCAGGAGGCCC	12.88	13.60
1	TAGGGTTAGGGTCCCCAGGT	TTGAA	2	AGGTTTCTGCCCATCCTTTC	GCCCAGGAAATCCTAAAGGT	0.044	0.039
2	GAGGGTTAGGGCCCCCAGGC	AGGGA	2	CCTACCAGCAGGAAAGGACA	CATCGTAACCGAAAGGTCCA	6.58	0.25
3	TAAGGTTCTGGGCCCCAGGC	AAGAA	3	CAGTGACTCACAGGGTCAGG	GGCGTTCCTATTTCACAAGC	0.052	0.046
4	AAGAGCTAGGGGCCCCAGGC	CTGAG	3	AAAAGGGGGTGGACTAGAGC	CACCAGGCCTGAGAGAGAAG	0.011	0.037
5	TATGTTTCGGGGCCCCAGGC	CGGAA	3	CACCTTCTGCATTCTGCCTA	TCCAGACCCTCAAAGACCAC	0.023	0.006
6	GAGGGGAAGGGGCCCCAGGC	TGGAG	3	GCAAAGACGGAAAGAGAAGC	CAGAGCCTTCAGAAATTCTCC	0.145	0.022
7	TAGGGGCAGGGGCACCAGGC	GGGGA	3	CCGTCTTGCTGTGTGACCTA	ATACGGACGCTCTGATCCTG	0.466	0.052
8	CCGGGTGAGTGGCCCCAGGC	CTGGG	4	CGACGTGAAGGAGAAATTCG	GCCAGTCGGAACACTCTGA	0.10	0.051
9	GAGGGTGAGTGGCCCCAGGG	CAGAA	4	AACCTGGAGTGGGATGACAG	CCACAGGGACTCTGAGGAGA	0.032	0.010
10	CAGGTTTAGGGGCTCCAGGA	CTGGG	4	TCTGTCCTCTGGGAGCTGAC	GCTTTGCAGACACCATCTCA	0.025	0.024
11	TGGGTTTAGGGGCCACAGGT	GGGAG	4	GGGCTCTGGCTTCTGAGAG	CTGGGTGCTCTCTACGTGGT	0.055	0.12
12	TGGGGTCAGGGGACCCAGGG	TGGGG	4	GGGGAGTGTTTTCCTTCCAT	GCCAGGGCTCACAGTTATTG	0.031	0.047
13	TAGGGTTAGGGGCCTGCAGC	CAGGG	4	CAGTCCTATGCTCGGGAGAG	GGGAACTGTAGCCTGTGGAG	0.015	0.024
14	TGGGGTGAGGGGCCCCGGCC	AGGAG	4	CAGAGGCTTCAGGAGGAAGG	TGGGGATATGCAACCCTTAG	0.13	0.15
15	GAGGATTAGGGTCACCAGGC	ATGAG	4	CTGGCAGGGGAAGTCAAATA	ATTCCGTCTGTCTGGAATGC	0.058	0.083
16	TGGGGCCAGGGGCCGCAGGC	AGGGG	4	CCCGTTCTCTCTCCTTCCTC	TGCACCAAGTAGCAGAGGTG	0.009	0.004
17	ACGGGTTAGGGGACACAGGC	CTGAG	4	CCTCTCTGAGCCCAGTGTTC	TCTTGTTCTCCACCCCTCAG	0.033	0.028
18	GAGGGGCAGGGGGCCCAGGC	TGGGG	4	GTCTGCTGGGATTCTGGGTA	CAGCTTTGTGGCTCTGGAAT	0.044	0.054
19	GAGCGTTGGGGGCCCCAGGA	CAGGA	4	CTCGTGAGCAACGGGACTAT	GTGGAAACACGGTGCTCTTT	0.066	0.062
20	TAGAGTTAGGAGACCCAGGA	ATGAG	4	CAACCAAGATCAGGCAACAA	AACTTGGTAAGTGCCCAGCA	0.12	0.066
21	TGGGGAGGGGGGCCCCAGGC	AGGGG	4	GGCCTCTGAAATAACGTTGG	CCCTGCTTTCTTCACTCCAG	0.043	0.057
22	AAGGGTTAGGGGCCCAAAGG	TAGAG	4	GGACCCTGGGAACATTTTGT	AAAGGGCAGAGGAAAGAAGG	0.046	0.066
23	GAGGCTGAGTGGCCCCAGGC	CTGAG	4	CCCAGTTTGAGGACAGTGGT	GGGCTTAGGGACTCAGGAGA	0.11	0.092

(Supp Table 10 Continued)

24	TCGGGTGTGGGG CTCCAGGC	CCGGG	4	CAAGAGAGGGAGG ATGCAAG	GCTGCTGAGGGA TGGAGTT	0.036	0.061
25	GAGGGTGAGTGG CCCCAGGA	CTGGG	4	CACAGACTCAGGC CATCTCA	GCAGTGAAAGAA GGCTAGATCC	0.084	0.031
26	TAGTGTTAGGAG CTCCAGGG	AAGGG	4	CCTACAGCCATTG GACCCTA	CGAAGGGCTCAA ACATCTTC	0.0030	0.0040
27	TAGGGTCAGGGG CTCAAGGG	ATGGG	4	GTCAGTGCTGACA CCTCACC	AGTGCCTCCTCT TCCCACTC	0.015	0.005
28	CAGGGATAGCAG CCCCAGGC	AGGGG	4	TGCTAGGGTGGGG AAATTCT	AAATCCAGCAGA GCAGCAAT	0.029	0.023
29	TAGGGGTAGGGG GGCCATGC	AGGGG	4	ACAGAAGGTAAGG GGGAAGG	TCTCTCTCTGCT GCACCTCA	0.074	0.058
30	TGGGGGTAGGGG TCCCAGGA	GAGAG	4	ATACCTGGGGGAA CTGCTCT	GTAGGCCACCTT GACCTCTG	0.015	0.015
31	CAGGCTTGGGGG CCCCAGGT	AGGGG	4	TCTGAGAACACCA GGAAGCA	TCTTGGCCTCCT CACATAGG	0.009	0.013

Appendix B Materials and Methods

Cell culture and transfection

Human embryonic kidney (HEK) cell line 293FT (Life Technologies) or mouse Neuro 2a (Sigma-Aldrich) cell line was maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10 % fetal bovine serum (HyClone), 2mM GlutaMAX (Life Technologies), 100 U/mL penicillin, and 100μg/mL streptomycin at 37 °C with 5 % CO_2 incubation.

Cells were seeded onto 24-well plates (Corning) at a density of 120,000 cells/well, 24 hours prior to transfection. Cells were transfected using Lipofectamine 2000 (Life Technologies) at 80-90 % confluency following the manufacturer's recommended protocol. A total of 500 ng Cas9 plasmid and 100 ng of U6-sgRNA PCR product was transfected.

Human embryonic stem cell line HUES62 (Harvard Stem Cell Institute core) was maintained in feeder-free conditions on GelTrex (Life Technologies) in mTesR medium (Stemcell Technologies) supplemented with 100 ug/ml Normocin (InvivoGen). HUES62 cells were transfected with Amaxa P3 Primary Cell 4-D Nucleofector Kit (Lonza) following the manufacturer's protocol.

SURVEYOR nuclease assay for genome modification

293FT and HUES62 cells were transfected with DNA as described above. Cells were incubated at 37 °C for 72 hours post-transfection prior to genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicentre) following the manufacturer's protocol. Briefly, pelleted cells were resuspended in QuickExtract solution and incubated at 65 °C for 15 minutes, 68 °C for 15 minutes, and 98 °C for 10 minutes.

The genomic region flanking the CRISPR target site for each gene was PCR amplified, and products were purified using QiaQuick Spin Column (Qiagen) following the manufacturer's protocol. 400 ng total of the purified PCR

products were mixed with 2ml 10X Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 20ml, and subjected to a re-annealing process to enable heteroduplex formation: 95 ºC for 10 min, 95 ºC to 85 ºC ramping at – 2 ºC/s, 85 ºC to 25 ºC at – 0.25 ºC/s, and 25 ºC hold for 1 minute. After re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (Transgenomics) following the manufacturer's recommended protocol, and analyzed on 4-20% Novex TBE poly-acrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 30 minutes and imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities. Indel percentage was determined by the formula, $100 \times (1 - (1 - (b + c) / (a + b + c))^{1/2})$, where a is the integrated intensity of the undigested PCR product, and b and c are the integrated intensities of each cleavage product.

Northern blot analysis of tracrRNA and sgRNA expression in human cells

Northern blots were performed as previously described (*1*). Briefly, RNAs were extracted using the mirPremier microRNA Isolation Kit (Sigma) and heated to 95 ºC for 5 min before loading on 8 % denaturing polyacrylamide gels (SequaGel, National Diagnostics). Afterwards, RNA was transferred to a pre-hybridized Hybond N+ membrane (GE Healthcare) and cross-linked with Stratagene UV Crosslinker (Stratagene). Probes were labeled with [gamma-32P] ATP (Perkin Elmer) with T4 polynucleotide kinase (New England Biolabs). After washing, membrane was exposed to phosphor screen for one hour and scanned with phosphorimager (Typhoon).

In vitro transcription and cleavage assay

Whole cell lysates from 293FT cells were prepared with lysis buffer (20 mM HEPES, 100 mM KCl, 5 mM MgCl2, 1 mM DTT, 5% glycerol, 0.1% Triton X-100) supplemented with Protease Inhibitor Cocktail (Roche). T7-driven

sgRNA was transcribed in vitro using custom oligos and HiScribe T7 In Vitro Transcription Kit (NEB), following the manufacturer's recommended protocol. To prepare methylated target sites, pUC19 plasmid was methylated by M.SssI and tested by digestion with HpaII. Unmethylated and successfully methylated pUC19 plasmids were linearized by NheI. The in vitro cleavage assay was performed as follows: for a 20 uL cleavage reaction, 10 uL of cell lysate was incubated with 2 uL cleavage buffer (100 mM HEPES, 500 mM KCl, 25 mM MgCl2, 5 mM DTT, 25% glycerol), 1 ug in vitro transcribed RNA, and 300 ng pUC19 plasmid DNA.

Bisulfite sequencing to assess DNA methylation status

Genomic DNA from 293FT cells was isolated with the DNeasy Blood & Tissue Kit (Qiagen) and bisulfite converted with EZ DNA Methylation-Lightning Kit (Zymo Research). Bisulfite PCR was conducted using KAPA2G Robust HotStart DNA Polymerase (KAPA Biosystems) with primers designed using the Bisulfite Primer Seeker (Zymo Research). Resulting PCR amplicons were gel-purified, digested with EcoRI and HindIII, and ligated into a pUC19 backbone prior to transformation. Individual clones were then Sanger sequenced to assess DNA methylation status.

Deep sequencing to assess targeting specificity

HEK 293FT cells were plated and transfected as described above, 72 hours prior to genomic DNA extraction. The genomic region flanking the CRISPR target site for each gene was amplified by a fusion PCR method to attach the Illumina P5 adapters as well as unique sample-specific barcodes to the target. PCR products were purified using EconoSpin 96-well Filter Plates (Epoch Life Sciences) following the manufacturer's recommended protocol.

Barcoded and purified DNA samples were quantified by Qubit 2.0 Fluorometer (Life Technologies) and pooled in an equimolar ratio. Sequencing libraries were then sequenced with the Illumina MiSeq Personal Sequencer (Life Technologies).

Sequencing data analysis and indel detection

MiSeq reads were filtered by requiring an average Phred quality (Q score) of at least 23, as well as perfect sequence matches to barcodes and amplicon forward primers. Reads from on- and off-target loci were analyzed by first performing Smith-Waterman alignments against amplicon sequences that included 50 nucleotides upstream and downstream of the target site (a total of 120 bp). Alignments, meanwhile, were analyzed for indels from 5 nucleotides upstream to 5 nucleotides downstream of the target site (a total of 30 bp). Analyzed target regions were discarded if part of their alignment fell outside the MiSeq read itself, or if matched base-pairs comprised less than 85% of their total length(*2*).

Negative controls for each sample provided a gauge for the inclusion or exclusion of indels as putative cutting events. For each sample, an indel was counted only if its quality score exceeded $\mu - \sigma$, where μ was the mean quality-score of the negative control corresponding to that sample and σ was the standard deviation of the same. This yielded whole target-region indel rates for both negative controls and their corresponding samples. Using the negative control's per-target-region-per-read error rate, q, the sample's observed indel count n, and its read-count R, a maximum-likelihood estimate for the fraction of reads having target-regions with true-indels, p, was derived by applying a binomial error model, as follows.

Letting the (unknown) number of reads in a sample having target regions incorrectly counted as having at least 1 indel be E, we can write (without making any assumptions about the number of true indels)

$$\text{Prob}(E|p) = \binom{R(1-p)}{E} q^E (1-q)^{R(1-p)-E}$$

since $R(1-p)$ is the number of reads having target-regions with no true indels. Meanwhile, because the number of reads observed to have indels is $n, n = E + Rp$, i.e. the number of reads having target-regions with errors but no true indels plus the number of reads whose target-regions correctly have indels. We can then re-write the above

$$\text{Prob}(E|p) = \text{Prob}(n = E + Rp|p) = \binom{R(1-p)}{n-Rp} q^{n-Rp}(1-q)^{R-n}$$

Taking all values of the frequency of target-regions with true-indels p to be equally probable a priori, $\text{Prob}(n|p) \propto \text{Prob}(p|n)$. The maximum-likelihood estimate (MLE) for the frequency of target regions with true-indels was therefore set as the value of p that maximized $\text{Prob}(n|p)$. This was evaluated numerically.

In order to place error bounds on the true-indel read frequencies in the sequencing libraries themselves, Wilson score intervals(3) were calculated for each sample, given the MLE-estimate for true-indel target-regions, Rp, and the number of reads R. Explicitly, the lower bound l and upper bound u were calculated as

$$l = \left(Rp + \frac{z^2}{2} - z\sqrt{Rp(1-p) + z^2/4}\right)/(R + z^2)$$

$$u = \left(Rp + \frac{z^2}{2} + z\sqrt{Rp(1-p) + z^2/4}\right)/(R + z^2)$$

where z, the standard score for the confidence required in normal distribution of variance 1, was set to 1.96, meaning a confidence of 95%.

Microinjection into mouse zygotes

Cas9 mRNA and sgRNA templates were amplified with T7 promoter sequence-conjugated primers. After gel purification, Cas9 and Cas9n were transcribed with mMESSAGE mMACHINE T7 Ultra Kit (Life technologies). sgRNAs were transcribed with MEGAshortscript T7 Kit (Life technologies). RNAs were purified by MEGAclear Kit (Life technologies) and frozen at -80 °C.

MII-stage oocytes were collected from 8-week old superovulated BDF1 females by injecting 7.5 I.U. of PMSG (Harbor, UCLA) and hCG (Millipore). They were transferred into HTF medium supplemented with 10 mg/ml bovine serum albumin (BSA; Sigma-Aldrich) and inseminated with capacitated sperm obtained from the caudal epididymides of adult C57BL/6 male mice. Six hours after fertilization, zygotes were injected with mRNAs and sgRNAs in M2 media (Millipore) using a Piezo impact-driven micromanipulator (Prime Tech Ltd., Ibaraki, Japan). The concentrations of Cas9 and Cas9n mRNAs and sgRNAs are described in the text. After microinjection, zygotes were cultured in KSOM (Millipore) in a humidified atmosphere of 5 % CO_2 and 95 % air at 37 °C.

Genome extraction from blastocyst embryos

Following in vitro culture of embryos for 6 days, the expanded blastocysts were washed with 0.01 % BSA in PBS and individually collected into 0.2 mL tubes. Five microliters of genome extraction solution (50 mM Tris-HCl, pH 8.0, 0.5 % Triton-X100, 1 mg/ml Proteinase K) were added and the samples were incubated in 65 °C for 3 hours followed by 95 °C for 10 min. Samples were then amplified for targeted deep sequencing as described above.

Western blot analysis

HEK 293FT cells were transfected and lysed in 1X RIPA buffer (Sigma-Aldrich) supplemented with Protease Inhibitor (Roche). The lysates were loaded onto Bolt 4–12% Bis-Tris Plus Gels (Invitrogen) and transferred to nitrocellulose membranes. The membranes were blocked in Tris-buffered saline containing 0.1% Tween-20 and 5% blocking agent (G-Biosciences). The membranes were probed with rabbit anti-FLAG (1:5,000, Abcam), HRP-conjugated anti-GAPDH (1:5,000 Cell Signaling Technology), and HRP-conjugated anti-rabbit (1:1,000) antibodies and visualized with a Gel Doc XR+ System (Bio-Rad).

Sample preparation for crystallography

The gene encoding full-length *S. pyogenes* Cas9 (residues 1–1368) was cloned between the *Nde*I and *Xho*I sites of the modified pCold-GST vector (TaKaRa), and mutations were introduced by a PCR-based method. The *S. pyogenes* Cas9 D10A/C80L/C574E/H840A mutant was expressed at 20°C in *Escherichia coli* Rosetta 2 (DE3) (Novagen), and was purified by chromatography on Ni-NTA Superflow resin (QIAGEN). The eluted protein was incubated overnight at 4°C with TEV protease to remove the His_6–GST-tag, and was further purified by chromatography on Ni-NTA, Mono S (GE Healthcare) and HiLoad Superdex 200 16/60 (GE Healthcare) columns. The SeMet-labeled Cas9 protein was expressed in *E. coli* B834 (DE3), and was purified using a similar protocol to that for the native protein. The 98-nt sgRNA was transcribed *in vitro* with T7 polymerase using a PCR-amplified DNA template, and was purified by 10% denaturing polyacrylamide gel electrophoresis. The 23-nt target DNA was purchased from Sigma-Aldrich. The purified Cas9 protein was mixed with sgRNA and DNA (molar ratio, 1:1.5:2.3), and then the complex was purified by gel filtration chromatography on a Superdex 200 Increase column (GE Healthcare), in a buffer consisting of 10 mM Tris-HCl, pH 8.0, 150 mM NaCl and 1 mM DTT.

Crystallography

The purified Cas9–sgRNA–DNA complex was crystallized at 20°C by the hanging-drop vapor diffusion method. Crystals were obtained by mixing 1 μl of complex solution ($A_{260\,nm}$, 15) and 1 μl of reservoir solution (13% PEG 3,350, 100 mM Tris-HCl, pH 8.0, 200 mM ammonium acetate and 100 mM NDSB-256). The SeMet-labeled protein was crystallized under conditions similar to those for the native protein. X-ray diffraction data were collected at 100 K on beamlines BL32XU and BL41XU at SPring-8 (Hyogo, Japan). The crystals were cryoprotected in reservoir solution supplemented with 25% ethylene glycol. X-ray diffraction data were processed using XDS (*4*). The structure was determined by the SAD method, using a 2.6 Å resolution data set from the

SeMet-labeled crystals. Forty of the potential 44 Se atoms were located using SHELXD (*5*) and autoSHARP (*6*). The initial phases were calculated using autoSHARP, and further improved by 2-fold NCS averaging using DM (*7*). The model was automatically built using PHENIX AutoSol (*8*), followed by manual model building using COOT (*9*) and structural refinement using PHENIX (*8*). The resulting model was further refined, using the 2.5 Å resolution native data set.

in vitro PAM screen and sgRNA prediction

Rho-independent transcriptional termination was predicted using the ARNold terminator search tool(*10, 11*). For the PAM library, a degenerate 7-bp sequence was cloned into a pUC19 vector. For each ortholog, the in vitro cleavage assay was carried out as above with 1 μg T7-transcribed sgRNA and 400ng pUC19 with degenerate PAM. Cleaved plasmids were linearized by *Nhe*I, gel extracted, and ligated with Illumina proprietary sequencing adaptors. Barcoded and purified DNA libraries were quantified by Quant-iT PicoGreen dsDNA Assay Kit or Qubit 2.0 Fluorometer (Life Technologies) and pooled in an equimolar ratio for sequencing using the Illumina MiSeq Personal Sequencer (Life Technologies). sgRNA secondary structure prediction was based on the Constraint Generation RNA folding model(*12*), and Sequence logos were generated using WebLogo(*13*).

AAV Production

For viral production, 293FT cells (Life Technologies) were maintained as recommended by the manufacturer in antibiotic-free media (DMEM, high glucose with GlutaMax and Sodium Pyruvate, supplemented with 10% FBS, and a final concentration of 10mM HEPES). For each vector, cells were grown in at least ten 15 cm tissue culture dishes and incubated until they reach around 70% - 80% confluence at 37°C and 5% CO_2. For transfection of virus production plasmids, PEI "Max" (Polysciences) was dissolved in water at 1 mg/mL and the pH of the solution was adjusted to 7.1.

For transfection, 8 ug of pAAV8 serotype packaging plasmid, 10 ug of pDF6 helper plasmid, and 6 ug of pAAV plasmid carrying the construct of interest were added to 1mL of serum-free DMEM. 125 uL of PEI "Max" solution was then added to the mixture. The resulting final transfection mixture was vortexed briefly and incubated at room temperature for 5 to 10 seconds. After incubation, the mixture was added to 20 mL of maintenance media, mix well, and applied to each dish to replace the old growth media. Cells were harvested between 48h and 72h post transfection. Cells were scraped from the dishes and pelleted by centrifugation. The AAV8 viral particle were then purified from the pellet according to previous published protocol(*14*).

Viruses were also produced by vector core facilities at University of Pennsylvania and Children's Hospital Boston, and titered by qPCR using a customized TaqMan probe against the SaCas9 transgene to match in house production.

Animal Injection and Processing

All mice were maintained at animal facility following IRB-approved protocols. AAV was delivered to at 8-10 week old C57/BL6 mice via tail vein injection. All dosages of AAV were adjusted to 100 uL or 200uL with sterile phosphate buffered serum, pH 7.4 (Gibco).

Tissue was harvested at the described time points post injection. Mice were anesthetized using Ketamine/Xylazine and subjected to transcardial perfusion with 30ml PBS. The median lobe of liver was removed and fixed in 4% paraformaldehyde for histological analysis, while the remaining lobes were sliced in small blocks of size less than $1x1x3mm^3$ and frozen at -80C for subsequent genomic DNA extraction, or immersed in RNALater (Ambion) for RNA extraction.

References

1. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819 (Feb 15, 2013).

2. P. Hsu *et al.*, DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827 (2013).

3. E. B. Wilson, Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc* **22**, 209 (Mar, 1927).

4. W. Kabsch, Xds. *Acta crystallographica. Section D, Biological crystallography* **66**, 125 (Feb, 2010).

5. G. M. Sheldrick, A short history of SHELX. *Acta crystallographica. Section A, Foundations of crystallography* **64**, 112 (Jan, 2008).

6. E. delaFortelle, G. Bricogne, Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol* **276**, 472 (1997).

7. K. Cowtan, 'dm': An automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 34 (1994).

8. P. D. Adams *et al.*, PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* **58**, 1948 (Nov, 2002).

9. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126 (Dec, 2004).

10. D. Gautheret, A. Lambert, Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of molecular biology* **313**, 1003 (Nov 9, 2001).

11. T. J. Macke *et al.*, RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic acids research* **29**, 4724 (Nov 15, 2001).

12. M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, K. P. Murphy, Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics* **23**, i19 (Jul 1, 2007).

13. G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, WebLogo: a sequence logo generator. *Genome Res* **14**, 1188 (Jun, 2004).

14. M. R. Veldwijk *et al.*, Development and optimization of a real-time quantitative PCR-based method for the titration of AAV-2 vector stocks. *Mol Ther* **6**, 272 (Aug, 2002).